1983

Use of transformed LANDSAT data in regression estimation of crop acreages

Hsien-Ming Hung
Iowa State University

Hung, Hsien-Ming

USE OF TRANSFORMED LANDSAT DATA IN REGRESSION ESTIMATION OF CROP ACREAGES

Iowa State University PH.D. 1983

Use of transformed LANDSAT data in regression estimation of crop acreages

by

Hsien-Ming Hung

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1983

TABLE OF CONTENTS

I. INTRODUCTION

Observations take one of two possible values with binary data. For instance, in a heart disease study, one possible binary response variable is the presence or absence of a particular heart disease. Mathematically, the two possible outcomes are both coded 1 and 0. If the binary response variable Y follows some probabilistic behavior, then the expected value of Y becomes the probability that the event '$Y = 1$' occurs. This probability is denoted by θ, where $\theta = p(Y = 1)$, and is often called the probability of success. In some cases one may be interested in studying the dependence of the probability of success on explanatory (auxiliary) variables. In fact, the expected value of Y conditional on $\underset{\sim}{X}$, denoted by $E(Y|\underset{\sim}{X})$, is known as the posterior probability of success conditional on $\underset{\sim}{X}$.

A linear logistic model is frequently used to model the relationship between θ and $\underset{\sim}{X}$. The logistic model is

$$\log[\theta(1-\theta)^{-1}] = \underset{\sim}{\beta}'\underset{\sim}{X},$$

where $\underset{\sim}{X}$ is the vector of explanatory variables and $\underset{\sim}{\beta}$ is a vector of unknown parameters. Alternatively, this logistic dependence can be postulated as

$$E(Y|\underset{\sim}{X}) = p(Y=1|\underset{\sim}{X}) = e^{\underset{\sim}{\beta}'\underset{\sim}{X}}(1 + e^{\underset{\sim}{\beta}'\underset{\sim}{X}})^{-1}.$$

The relation

$$E(Y|\underset{\sim}{X}) = p(Y=1|\underset{\sim}{X})$$

holds exactly provided that Y is a zero-one variable. In the logistic model, the parameter vector $\underset{\sim}{\beta}$ can be estimated using the maximum likelihood method when independent observations are available. The true posterior probability $p(Y=1|\underset{\sim}{X})$ can be estimated by substituting the estimate $\hat{\underset{\sim}{\beta}}$ for the parameter vector $\underset{\sim}{\beta}$.

In some types of surveys, the sampling unit consists of a group, or cluster, of smaller units that are called elements. Soil sampling and surveys of farming are examples of surveys where cluster sampling is frequently used. The quantity to be estimated is often the total number of elements in the population of N clusters that fall into some defined class C. The variable Y_{ij} is defined as 1 if the j-th element of the i-th cluster is in C, and $Y_{ij} = 0$ if it is not in C. The quantity of interest in this case is $Y_T = \Sigma_{i=1}^{N} \Sigma_{j=1}^{M_i} Y_{ij}$, where M_i is the number of elements in the i-th cluster. In single-stage simple random cluster sampling, an unbiased estimator of Y_T is the sample-mean estimator $\hat{Y}_T$, where $\hat{Y}_T = n^{-1} N \Sigma_{i=1}^{n} \Sigma_{j=1}^{M_i} y_{ij}$ and y_{ij} is the Y-value of the j-th element in the i-th sampled cluster.

When data from auxiliary variables $\underset{\sim}{X}$ are available, regression estimation techniques can be applied to improve estimation. The true posterior probability is the best possible auxiliary variable for each

element because it is the expected value conditional on $\underset{\sim}{X}$ of the binary variable Y. With cluster sampling, one possible auxiliary variable for each cluster is the sum of posterior probabilities, where the sum is over all the elements in the cluster. The primary variable is the sum of the Y values over all the elements in the cluster. One question is whether or not the probability sum is the best possible auxiliary variable for each cluster. Since the true posterior probability cannot be observed, it has to be estimated from the sample. The effect of the estimation of the posterior probability on the variance of the resulting regression estimator is another problem which will be examined.

The literature on discriminant analysis, regression analysis for complex surveys, and LANDSAT crop estimation is reviewed in Chapter II. In Chapter III, the effect of estimating the posterior probability on the variance of the resulting regression estimator will be investigated. This theoretical framework is also extended to the general case where auxiliary variates must be estimated in regression estimation. In Chapter IV, the performance of the estimated posterior probability will be compared to that of classification functions by using LANDSAT crop data. Finally, we present various methods of estimating the variance of the regression estimator, when the regression estimator is constructed by using estimated probability sum as an auxiliary variable.

II. LITERATURE REVIEW

A. Discriminant Analysis

Theories and methods of classification and discrimination have attracted many researchers from different disciplines, and a large body of literature is available in this area. One good bibliography on discriminant analysis was published by Cacoullos (1973). Moreover, results are available in the well-known textbooks by Anderson (1958) and by Rao (1952). The usual classification problem can be formulated as follows: Suppose an individual is an observation from one of several populations $\pi_1,\ldots, \pi_m$. The classification of an observation into one of the populations depends on the vector of measurements $\underset{\sim}{x} = (x_1, x_2,\ldots, x_p)'$ for that individual. We wish to divide the p-dimensional space of observations into m mutually exclusive regions $R_1,\ldots, R_m$ so that the individual is said to come from population π_i if the observation falls into R_i. Let $c(j|i)$ be the cost of misclassifying an observation from π_i as coming from π_j. Let the probability that an observation comes from population π_i be q_i and the density of population π_i be $p_i(\underset{\sim}{x})$, $i = 1,2,\ldots, m$. As shown in Anderson (1958), the regions of classifications, $R_1,\ldots, R_m$, that minimize the expected cost are defined by assigning an individual with $\underset{\sim}{x}$ to R_j if

$$\sum_{\substack{i=1\\ i\neq j}}^{m} q_i p_i(\underset{\sim}{x}) c(j|i) < \sum_{\substack{i=1\\ i\neq k}}^{m} q_i p_i(\underset{\sim}{x}) c(k|i) , \quad (k = 1,2,\ldots, m,\ k \neq j) .$$

In the case of two populations, the regions of classification, R_1 and R_2, that minimize the expected cost are given by

$$R_1 : \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} \geq \frac{c(1|2)q_2}{c(2|1)q_1} ,$$

$$R_2 : \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} < \frac{c(1|2)q_2}{c(2|1)q_1} .$$

Note that when $c(i|j)$ for all i,j are equal to one, the expected loss will be the probability of misclassification. In the case of two multivariate normal population with different means μ_1, μ_2 and common covariance matrix $\underset{\sim}{\Sigma}$,

$$\frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} = \underset{\sim}{x}'\underset{\sim}{\Sigma}^{-1}(\mu_1 - \mu_2) - \tfrac{1}{2}(\mu_1 + \mu_2)'\underset{\sim}{\Sigma}^{-1}(\mu_1 - \mu_2) . \quad (2.A.1)$$

The first term on the right is the well-known discriminant function which is a linear function of the components of $\underset{\sim}{x}$. When $\underset{\sim}{x}$ comes from $N(\mu_1, \underset{\sim}{\Sigma})$, the discriminant function U, where

$$U = \underset{\sim}{x}'\underset{\sim}{\Sigma}^{-1}(\mu_1 - \mu_2) - \tfrac{1}{2}(\mu_1 + \mu_2)'\underset{\sim}{\Sigma}^{-1}(\mu_1 - \mu_2) ,$$

has a normal distribution $N(\frac{1}{2}\Delta^2, \Delta^2)$, where Δ^2 is the Malahanobis distance,

$$\Delta^2 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) .$$

Similarly, U has a normal distribution $N(-\frac{1}{2}\Delta^2, \Delta^2)$ when $\underset{\sim}{x}$ comes from $N(\underset{\sim}{\mu}_2, \underset{\sim}{\Sigma})$. The probability of misclassification if the observation is from π_1 is $\Phi(-\frac{1}{2}\Delta)$, where $\Phi(z)$ is the standard normal distribution function defined by

$$\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-\frac{1}{2}} e^{-x^2/2} \, dx .$$

The probability of misclassification if the observation is from π_2 is also $\Phi(-\frac{1}{2}\Delta)$.

For the case where the population parameters are not known, a sample from each of two normal populations is needed to estimate the parameters $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \underset{\sim}{\Sigma}$. Assume random samples of size N_1 and N_2 have been drawn independently from the two p-dimensional multivariate normal populations, respectively. These samples are often called initial samples. Let $\bar{\underset{\sim}{x}}_1$ and $\bar{\underset{\sim}{x}}_2$ be the respective sample mean vectors, and let $\underset{\sim}{S}$ be the pooled estimator of $\underset{\sim}{\Sigma}$. The discriminant function suggested by Fisher (1936) is

$$\underset{\sim}{x}'\underset{\sim}{S}^{-1}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) .$$

This discriminant function is obtained using a linear combination of the observations and choosing the coefficients to maximize the ratio of the difference of the means of the linear combination in the two sample groups. Anderson (1951) proposed the discriminant function W by replacing the parameters in (2.A.1) by the estimates. The limiting distribution of W, was considered by Wald (1944) and Anderson (1958). As $N_1 \to \infty$ and $N_2 \to \infty$, the limiting distribution of W is normal with variance Δ^2 and mean $\frac{1}{2}\Delta^2$ if $\underset{\sim}{x}$ is from $N(\underset{\sim}{\mu}_1, \underset{\sim}{\Sigma})$ and mean $-\frac{1}{2}\Delta^2$ if $\underset{\sim}{x}$ is from $N(\underset{\sim}{\mu}_2, \underset{\sim}{\Sigma})$. For the Studentized W, Anderson (1973) proved that $(W - \frac{1}{2}\hat{\Delta}^2)/\hat{\Delta}$ and $(W + \frac{1}{2}\hat{\Delta}^2)/\hat{\Delta}$, where

$$\hat{\Delta}^2 = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) ,$$

have a standard normal as the limiting distributions when $\underset{\sim}{x}$ comes from π_1 and $\underset{\sim}{x}$ comes from π_2, respectively.

Estimation of misclassification probabilities (error rates), has received considerable attention. The simplest way to estimate error rates is to use the data points in the initial samples that are misclassified. The error rate calculated in this way is known as the apparent error rate. Unfortunately, this error rate leads to an optimistic result. In other words, it underestimates the true error rate which is the expected error rate of the classification procedure in future samples. McLachlan (1976) has derived the asymptotic bias of the apparent error rate for the case of two multivariate normal

populations. McLachlan has also shown that the average apparent error rate is asymptotically less than the average actual error rate. Hills (1966) showed that the expected actual error rate is less than $\Phi(-\Delta^2/2)$, and $\Phi(-\Delta^2/2)$ is less than the actual error rate. The error rate $\Phi(-\Delta^2/2)$ is often called the optimum error rate. The unconditional mean of the actual error rate is given by Lachenbruch (1968). Several other estimators for the expected actual error rate are proposed by Okamoto (1963), Lachenbruch and Mickey (1968), Anderson (1973a, 1973b), and McLachlan (1974a, 1974b, 1974c). Most of the estimators are obtained by adjusting the bias of $\Phi(-\hat{\Delta}^2/2)$. The estimator $\Phi(-\hat{\Delta}^2/2)$ generally performs poorest, while the Okamoto procedure with a special estimate of Δ^2 is satisfactory. McLachlan (1974a) suggested an asymptotically unbiased technique for estimating the actual error rate. This technique is to use the asymptotic expansion of the actual error rate and to adjust for the biases of first order and second order with respect to $(N_1^{-1}, N_2^{-1}, N^{-1})$, where $N = N_1 + N_2 - 2$. An asymptotic unbiased estimator of the average actual error rate was derived by McLachlan (1974c). The asymptotic mean square error (AMSE) was considered as a criterion for comparing several estimators by McLachlan (1974b). The AMSE of a given estimator, Q_t, is obtained by expanding the expectation of $(P_1 - Q_t)^2$ over the joint distribution of $(\bar{\underset{\sim}{x}}_1, \bar{\underset{\sim}{x}}_2, \underset{\sim}{S})$, where P_1 is the actual error rate. The terms of the third order with respect to $(N_1^{-1}, N_2^{-1}, N^{-1})$ are neglected. McLachlan (1974b) concluded that the relative superiority as

determined on the AMSE criterion agrees with the relative superiority on the basis of the absolute distance between the estimated and the true value of P_1 . An extensive bibliography of error rate estimation has been published by Toussaint (1974).

The equal covariance assumption in normal populations is rarely satisfied; however, most of the early work in classification theory was based on an equal covariance assumption. When the covariance matrices are quite different, the optimal rule in the sense of minimizing the error rates assigns an observation to π_1 if

$$Q(\underset{\sim}{x}) = \ell n \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} > \ell n \frac{1 - q_1}{q_1} .$$

The classification function $Q(\underset{\sim}{x})$ is quadratic in $\underset{\sim}{x}$ since $\underset{\sim}{\Sigma}_1^{-1} - \underset{\sim}{\Sigma}_2^{-1}$ does not vanish:

$$Q(\underset{\sim}{x}) = C_0 - 1/2 \, [\underset{\sim}{x}'(\underset{\sim}{\Sigma}_1^{-1} - \underset{\sim}{\Sigma}_2^{-1})\underset{\sim}{x} - 2\underset{\sim}{x}'(\underset{\sim}{\Sigma}_1^{-1} \underset{\sim}{\mu}_1 - \underset{\sim}{\Sigma}_2^{-1} \underset{\sim}{\mu}_2)] ,$$

where

$$C_0 = 1/2 \, \ell n(|\underset{\sim}{\Sigma}_2| / |\underset{\sim}{\Sigma}_1|) .$$

Gilbert (1969) undertook an investigation of the effect of unequal variance-covariance matrices on Fisher's linear discriminant function, when the parameters of the two normal populations are known. The

optimal probability of misclassification was calculated using the central chi-square to approximate the exact probability. Marks and Dunn (1974) conducted a Monte Carlo study to compare the performance of Fisher's linear discriminant function and the quadratic discriminant function when the two multivariate normally distributed populations have unequal covariance matrices. The results indicated that for small samples the quadratic discriminant function performs worse than the linear discriminant function when covariances are nearly equal and the dimension of $\underset{\sim}{x}$ is large. A Monte Carlo study conducted by Wahl and Kronmal (1977) suggested that sample size is an important consideration in deciding to use the quadratic discriminant function in multivariate normal situations. When the dimension of $\underset{\sim}{x}$ and the differences between two population covariances are large, the quadratic discriminant function performs much better than Fisher's linear discriminate function provided the sample size is sufficient. The comparisons were made based on the probabilities of misclassifications.

Consideration has been given by several statisticians to the situation where some or all of the observations are qualitative. Logistic discrimination for two populations was proposed by Cox (1966) and Day and Kerridge (1967) and has been further developed by Anderson (1972). For the case of m populations, $\pi_1, \ldots, \pi_m$, the logistic discriminant functions depend on the posterior probability that an individual with the vector $\underset{\sim}{x}$ is from π_i. This probability is given by

$$p(\pi_i|\underset{\sim}{x}) = \frac{p_i(\underset{\sim}{x})q_i}{\sum_{j=1}^{m} p_j(\underset{\sim}{x})q_j} . \qquad (2.A.2)$$

If $p_i(\underset{\sim}{x})$ is multivariate normal with mean $\underset{\sim}{\mu}_i$ and covariance matrix $\underset{\sim}{\Sigma}$, equation (2.A.2) can be written as

$$p(\pi_i|\underset{\sim}{x}) = \exp(\alpha_{0i} + \underset{\sim}{\beta}_i'\underset{\sim}{x})p(\pi_m|\underset{\sim}{x}) , \quad i = 1,\ldots, m-1 ,$$

$$p(\pi_m|\underset{\sim}{x}) = \frac{1}{1 + \sum_{i=1}^{m-1} \exp(\alpha_{0i} + \underset{\sim}{\beta}_i'\underset{\sim}{x})} .$$

For $m = 2$, we can write (2.A.2) as

$$p(\pi_1|\underset{\sim}{x}) = \exp(\alpha_0 + \underset{\sim}{\beta}'\underset{\sim}{x})p(\pi_2|\underset{\sim}{x}) ,$$

$$p(\pi_2|\underset{\sim}{x}) = \frac{1}{1 + \exp(\alpha_0 + \underset{\sim}{\beta}'\underset{\sim}{x})} ,$$

where

$$\underset{\sim}{\beta} = \underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) ,$$

$$\alpha_0 = -\tfrac{1}{2}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) + \ell n[q_1/(1 - q_1)] .$$

The coefficients α_0 and $\underset{\sim}{\beta}$ are usually estimated by maximum likelihood. In discriminant problems, samples are most often taken from each population separately. Anderson (1972) noted that if the populations are sampled separately, the only coefficient that is changed is α_0.

Duda and Hart (1973) give a systematic account of major topics in pattern recognition. In their setup, samples $\underset{\sim}{x}$ are assumed to be obtained by selecting a state of nature ω_j with probability $p(\omega_j)$, and then independently selecting $\underset{\sim}{x}$ according to the probability law $p(\underset{\sim}{x}|\omega_j)$. Suppose that $p(\underset{\sim}{x}|\omega_j) \sim N(\underset{\sim}{\mu}_j, \underset{\sim}{\Sigma}_j)$ for all m classes. Let n samples $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be drawn with the labels $\ell_1, \ldots, \ell_n$, respectively; i.e., $\ell_k = i$ if the state of nature for $\underset{\sim}{x}_k$ was ω_i. The maximum likelihood estimators for $\underset{\sim}{\mu}_i$ and Σ are given by

$$\hat{\underset{\sim}{\mu}}_i = n_i^{-1} \left[\sum_{\ell_k = i} \underset{\sim}{x}_k \right]$$

and

$$\hat{\underset{\sim}{\Sigma}}_i = [n_i - 1]^{-1} \sum_{k=1}^{n} (\underset{\sim}{x}_k - \hat{\underset{\sim}{\mu}}_{\ell_k})(\underset{\sim}{x}_k - \hat{\underset{\sim}{\mu}}_{\ell_k})' ,$$

provided that the number of samples n_i with label i is greater than one. When $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_k$ are unlabeled samples and $p(\underset{\sim}{x}|\omega_j) \sim N(\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma}_i)$ for all m classes, the maximum likelihood principle may yield useless singular solutions if no constraints are

placed on the covariance matrix. But the local-maximum-likelihood estimators $\hat{\mu}_i$, $\hat{\Sigma}_i$ and $\hat{p}(\omega_i)$ will satisfy

$$\hat{p}(\omega_i) = n^{-1} \sum_{k=1}^{n} \hat{p}(\omega_i | \underset{\sim}{x}_k) ,$$

$$\hat{\mu}_i = [\sum_{k=1}^{n} \hat{p}(\omega_i | \underset{\sim}{x}_k)]^{-1} [\sum_{k=1}^{n} \hat{p}(\omega_i | \underset{\sim}{x}_k) \underset{\sim}{x}_k] ,$$

$$\hat{\Sigma}_i = [\sum_{k=1}^{n} \hat{p}(\omega_i | \underset{\sim}{x}_k)]^{-1} [\sum_{k=1}^{n} \hat{p}(\omega_i | \underset{\sim}{x}_k)(\underset{\sim}{x}_k - \hat{\mu}_i)(\underset{\sim}{x}_k - \hat{\mu}_i)'] ,$$

where

$$\hat{p}(\underset{\sim}{\omega}_i | \underset{\sim}{x}_k) = \{ \sum_{j=1}^{m} |\hat{\Sigma}_j|^{-1/2} \exp[-1/2 (\underset{\sim}{x}_k - \hat{\mu}_j)' \hat{\Sigma}_j^{-1} (\underset{\sim}{x}_k - \hat{\mu}_j)] \hat{p}(\underset{\sim}{\omega}_j) \}^{-1}$$

$$\times |\hat{\Sigma}_i|^{-1/2} \exp[-1/2 (\underset{\sim}{x}_k - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} (\underset{\sim}{x}_k - \hat{\mu}_i)] \hat{p}(\omega_i) .$$

Another way to approach the problem of parameter estimation is the Bayesian estimation procedure. This method views the parameters as random variables having some known a priori distribution. It was noted, however, that the results obtained by maximum likelihood estimation and Bayesian estimation are frequently nearly identical.

If the forms of the class-conditional probability density functions $p(\underset{\sim}{x} | \pi_i)$ are unknown, then nonparametric methods can be used to estimate these probability density functions. Hand (1981) summarized four major types of nonparametric probability density function estimators: the

histogram method, the kernel method, the k-nearest-neighbor method, and the series expansion method. Each of these methods has different advantages and disadvantages. In the histogram procedure, the whole space is partitioned into disjoint cells of equal volume. The probability density function is estimated by the proportion of sample points falling in each cell. Discontinuities at the edges of the cells and a sudden drop to zero outside the boundary cells are two main problems associated with the histogram method.

Kernel estimators are developed based on the fact that density functions are derivatives of cumulative distribution functions. In one dimension, let $\nu(x|\pi_i)$ to be the number of class i sample points with values less than or equal to x. Let N_i be the total sample size for class i. An estimator of $p(x|\pi_i)$ defined as being an approximation to the derivative of the estimated cumulative distribution of x is given by

$$\hat{p}(x|\pi_i) = \frac{\nu(x + h|\pi_i) - \nu(x - h|\pi_i)}{2h\ N_i},$$

where h is an arbitrarily chosen positive constant. If $\{x_1, \ldots, x_{N_i}\}$ is the sample, then the natural generalization will lead to the derivation of $\hat{p}(x|\pi_i)$ as

$$\hat{p}(x|\pi_i) = (N_i h)^{-1} \sum_{t=1}^{N_i} K\left(\frac{x - x_t}{h}\right) , \qquad (2.A.3)$$

where

$$K(z) = \begin{cases} 0 & \text{if } |z| > 1 \\ 1/2 & \text{otherwise} \end{cases} .$$

In effect, each point in the closed interval $[x - h, x + h]$ contributes equally to the estimation of $p(x|\pi_i)$. Such a weighting has been generalized by using smoothness properties. The extension of (2.A.3) to higher dimensions is straightforward. The general form of kernel estimators is given by

$$\hat{p}(\underset{\sim}{x}|\pi_i) = (N_i)^{-1} \sum_{t=1}^{N_i} K(\underset{\sim}{x} - \underset{\sim}{x}_t) ,$$

where $K(\underset{\sim}{z})$ is the so called kernel function satisfying

(i) $K(\underset{\sim}{z}) \geq 0$ and

(ii) $\int K(\underset{\sim}{z})\, d\underset{\sim}{z} = 1$.

The remaining matter is to choose the kernel function, which is often the problem associated with the kernel method.

In the k-nearest-neighbor methods, a cell is centered about $\underset{\sim}{x}$ and it is allowed to grow until k samples are captured, where k is some specified function of N and $N = \Sigma_{i=1}^{m} N_i$. If among these k points there occur k_i points from class π_i , then a k-nearest-neighbor estimator of $p(\underset{\sim}{x}|\pi_i)$ is given by

$$\hat{p}(\underset{\sim}{x}|\pi_i) = N_i^{-1} V^{-1} k_i ,$$

where V is the volume of the cell centered at $\underset{\sim}{x}$. The estimate for the posterior probability $p(\pi_i|\underset{\sim}{x})$ is

$$\hat{p}(\pi_i|\underset{\sim}{x}) = k^{-1} k_i .$$

This leads immediately to the classification rule: classify $\underset{\sim}{x}$ as belonging to class i if $k_i = \max\{k_1, \ldots, k_m\}$. This is known as the k-nearest-neighbor classification rule. A disadvantage common to both the kernel and the k-nearest-neighbor methods is that all of the sample points need to be retained. In other words, the distances from $\underset{\sim}{x}$ to all of the sample points must be determined.

The series expansion method approximates the kernel function by a finite series expansion. More explicitly, let $\{\phi_i\}$ be the set of orthonormal basis functions so that

$$\int \phi_i(\underset{\sim}{x})\phi_j(\underset{\sim}{x})d\underset{\sim}{x} = \delta_{ij} ,$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Then the series expansion estimator of $p(\underset{\sim}{x}|\pi_i)$ is

$$\hat{p}(\underset{\sim}{x}|\pi_i) = \sum_{j=1}^{s} \{N_i^{-1} \sum_{k=1}^{N_i} \phi_i(\underset{\sim}{x}_k)\}\phi_j(\underset{\sim}{x}) ,$$

where s is an arbitrary positive integer. The series method has an advantage that the memory and data storage can be saved. It should be noted, however, that it may require a large number of terms s to make the series expansion accurate in the region of interest.

B. Regression Estimation in Complex Surveys

Most standard statistical methods have been developed on the assumption of independent observations. This assumption of independence is reasonable, especially when data are collected in controlled experiments. However, much research work is carried out with complex sample designs, especially in social, health, economic and agricultural studies. Almost all large surveys are cluster or multi-stage samples for economic reasons.

Cluster samples are characterized by units selected in groups called clusters. These clusters, also called primary sampling units, consist of smaller units that are called elements or subunits. If only a subset of the subunits in each cluster is observed, the sample is called a multi-stage cluster sample. Let the population contain N clusters of size M_i, $i = 1,2,\ldots, N$, with $M_0 = \Sigma_{i=1}^{N} M_i$ elements in total. Associated with each of these elements is a vector of $p + 1$

values $(Y_{ij}, \underset{\sim}{X}_{ij})$, where $\underset{\sim}{X}_{ij} = (X_{1ij}, X_{2ij}, \ldots, X_{pij})'$, $j = 1,2,\ldots, M_i$, $i = 1,2,\ldots, N$. A sample of n clusters is drawn from this population by a probabilistic selection procedure which is called a sampling design. For convenience, let the first n population clusters be sampled. Then a sample of n vectors from the population of N vectors is given by $S = \{(Y_{ij}, \underset{\sim}{X}'_{ij}) \mid j = 1,2,\ldots, M_i$ $i = 1,2,\ldots, n\}$. The variable Y is the characteristic of interest, and $\underset{\approx}{X}$ serves as auxiliary information.

The classical theory of regression assumes a linear relationship among the variables. The model is

$$Y_{ij} = \underset{\sim}{X}'_{ij}\underset{\sim}{\beta} + \underset{\sim}{e}_{ij} ,$$

where the unknown parameter vector is $\underset{\sim}{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\underset{\sim}{e}_{ij}$ is the error. In addition, the following assumptions are often made:

(i) $E(e_{ij} \mid \underset{\sim}{X}_{ij}) = 0$, for all i,j ;

(ii) $E(e_{ij}^2 \mid \underset{\sim}{X}_{ij}) = \sigma^2$, for all i,j ;

(iii) $E(e_{ij}e_{i'j'} \mid \underset{\sim}{X}_{ij}, \underset{\sim}{X}_{i'j'}) = 0$, for all $(i,j) \neq (i',j')$;

(iv) Normality for the e_{ij} .

Standard least-squares method can be used to find $\hat{\underset{\sim}{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ such that

$$\sum_{i=1}^{n} \sum_{j=1}^{M_i} (Y_{ij} - \sum_{\ell=1}^{p} \hat{\beta}_\ell X_{\ell ij})^2$$

is minimized. With the four assumptions, the least squares procedure yields many desirable results. For example, $\hat{\underset{\sim}{\beta}}$ is the best linear unbiased estimator.

In survey sampling, most target populations are finite populations rather than infinite populations. Also, a complex selection design tends to introduce correlation between elements so that assumption (iii) fails to hold. For instance, in cluster sampling, clusters often exhibit positive intracluster correlation, the principal effect of which is to increase the variances of the estimators of the population mean as compared to the variance achieved under simple random sampling. However, even without assumption about the relation between Y and $\underset{\sim}{X}$, the regression estimation approach for population means (or population totals) has been well-developed and can be found in most textbooks. Large-sample results are available for the survey sampling model, but very little is known for small samples.

Several statisticians have proposed a model in which the finite population is a sample from an infinite superpopulation. The finite population total Y_T is regarded as a fixed quantity in classical sampling theory; whereas, under the superpopulation model Y_T is a random variable. Assumptions (i) - (iv) are often made for the superpopulation. This approach provides some information on the

efficiency of the estimators in moderate or small samples and on sample size requirements for the practical use of large-sample results.

Several contributions to regression analysis in cluster samples will be reviewed and summarized. First, let us introduce some additional notation. Denote by $\bar{\underset{\sim}{W}}_{(p)}$ and $\bar{\underset{\sim}{W}}_{(s)}$ the population mean and the sample mean respectively for a random vector $\underset{\sim}{W}$. Let $\underset{\sim}{W}_{i.}$ be the i-th cluster total of $\underset{\sim}{W}$, and let $\bar{\underset{\sim}{W}}_{i.}$ be the i-th cluster mean of $\underset{\sim}{W}$, $i = 1,2,\ldots, N$.

Cochran (1942) discussed the use of cluster sizes in making estimates from a sample. An infinite population was assumed, and $Y_{i.}$ was assumed to be linearly related to the cluster sizes M_i. Thus, his model is

$$Y_{i.} = \alpha + \beta M_i + e_i ,$$

where e has zero mean and constant variance. The regression estimator for the population total Y_T is

$$\hat{Y}_T = N[\bar{Y}_{(s)} + \hat{\beta}(\bar{M}_{(p)} - \bar{M}_{(s)})] ,$$

where $\hat{\beta}$ is the sample regression coefficient

$$\hat{\beta} = [\sum_{i=1}^{n} (M_i - \bar{M}_{(s)})^2]^{-1}[\sum_{i=1}^{n} (Y_{i.} - \bar{Y}_{(s)})(M_i - \bar{M}_{(s)})] .$$

The sampling variance of $\hat{Y}_T$, given that the M's are fixed, is

$$V(\hat{Y}_T|M_1,\ldots, M_N) = N^2\ \sigma_y^2(1 - \rho^2)\left[\frac{1}{n} + \frac{(\bar{M}_{(p)} - M_{(s)})^2}{\sum_{i=1}^{n} (M_i - \bar{M}_{(s)})^2}\right],$$

where ρ is the population correlation coefficient between Y and M . A sample estimator of this variance is obtained by substituting the residual mean square error s_d^2 from the sample regression for $\sigma_y^2(1 - \rho^2)$. As summarized in Cochran (1977), under random sampling, to the terms of order n^{-2} ,

$$E[V(\hat{Y}_t|M_1,\ldots, M_N)] = \frac{N^2\ \sigma_y^2(1 - \rho^2)}{n}\left(1 + \frac{1}{n-3} + \frac{2G_1^2}{n^2}\right),$$

where $G_1 = k_{3M}/\sigma_M^3$ is Fisher's measure of relative skewness of the distribution of M .

In the case where the straight line passes through origin, the regression estimator of the population total becomes

$$\tilde{Y}_T = N\ \bar{M}_{(p)}\left(\sum_{i=1}^{n} M_i^2\right)^{-1}\left(\sum_{i=1}^{n} M_i Y_{i.}\right)$$

$$= M_0\left(\sum_{i=1}^{n} M_i^2\right)^{-1}\left(\sum_{i=1}^{n} M_i Y_{i.}\right),$$

where M_0 is the population total of cluster sizes. The variance of $\tilde{Y}_T$, given that the M's are fixed, is

$$V(\tilde{Y}_T|M_1,\ldots M_N) = M_0^2\ \sigma_y^2(1-\rho^2)/\sum_{i=1}^{n} M_i^2 .$$

The expected variance of $\tilde{Y}_T$ under random sampling is approximately

$$E[V(\hat{Y}_T|M_1,\ldots, M_N)] = N^2\ \sigma_y^2(1-\rho^2)[n(1+C_M)]^{-1}\left[1+\frac{2\ C_M(2+C_M)}{n(1+C_M)^2}\right],$$

where $C_M = \sigma_M^2/\bar{M}_{(p)}^2$ is the square of the coefficient of variation of M, provided that the distribution of M is not far from normal.

Ignoring the information in the cluster sizes, a possible estimate of Y_T is $\dot{Y}_T = N\ \bar{Y}_{(s)}$. Under random sampling, the expected variance of $\dot{Y}_T$ is $N^2n^{-1}\sigma_y^2$. Omitting terms of order n^{-2} and n^{-3}, comparisons among these three estimators are as follows:

$$\frac{E[V(\hat{Y}_T|M_1,\ldots, M_N)]}{E[V(\dot{Y}_T)]} = 1-\rho^2 ,$$

$$\frac{E[V(\tilde{Y}_T|M_1,\ldots, M_N)]}{E[V(\dot{Y}_T)]} = \frac{1-\rho^2}{1+C_M} ,$$

and

$$\frac{E[V(\tilde{Y}_T|M_1,\ldots,M_N)]}{E[V(\hat{Y}_T|M_1,\ldots,M_N)]} = \frac{1}{1 + C_M},$$

where $E[V(\dot{Y}_T)]$ is defined to be the expected variance of $\dot{Y}_T$. These results indicate that in large samples $\hat{Y}_T$ and $\tilde{Y}_T$ are never less accurate on the average than $\dot{Y}_T$. Also, $\tilde{Y}_T$ is more accurate than $\hat{Y}_T$ when the true regression line is a straight line passing through the origin. However, if the line passes through a point $(0, \alpha)$, the estimate $\tilde{Y}_T$ is biased with the bias tending to a constant value $N\alpha C_M(1 + C_M)^{-1}$ in large samples. The average mean square error of $\tilde{Y}_T$ is

$$E[MSE(\tilde{Y}_T)] = \frac{N^2\alpha^2 C_M}{(1 + C_M)^2} + \frac{N^2\sigma_y^2(1 - \rho^2)}{n(1 + C_M)}.$$

Note that the component arising from the bias does not decrease as the number of clusters, n, increases. Both $\hat{Y}_T$ and $\dot{Y}_T$ will be more accurate than $\tilde{Y}_T$ for large n. Cochran did not recommend $\tilde{Y}_T$ as a good estimator unless one is certain that the true line passes through the origin.

Cochran (1942) also investigated how the linear regression estimator is affected when the population regression model is nonlinear. He described this situation by giving a model of the form

$$y_{i.} = \alpha + \beta M_i + \xi_i + e_i ,$$

where ξ is nonlinear function of M with zero mean, and e, distributed with zero mean and unit variance, is independent of M. The error of $\hat{Y}_T$ is then given by

$$\hat{Y}_T - Y_T = N\left[(\bar{\xi}_{(s)} + \bar{e}_{(s)}) + (\bar{M}_{(p)} - \bar{M}_{(s)}) \frac{\sum_{i=1}^{n} (\xi_i + e_i)(M_i - \bar{M}_{(s)})}{\sum_{i=1}^{n} (M_i - \bar{M}_{(s)})^2}\right].$$

It follows that $\hat{Y}_T$ is biased and the bias is of order n^{-1}. The expected sampling variance of $\hat{Y}_T$ is also changed, but only the terms of order n^{-1} or n^{-2} are different. Ultimately, the bias becomes negligible relative to the standard error as the number of sampled clusters becomes sufficiently large. Also, the large-sample efficiency of $\hat{Y}_T$ to the sample-mean estimate $\dot{Y}_T$ remains $(1 - \rho^2)$. On the other hand, the term $n^{-1} \sigma_y^2(1 - \rho^2)$ becomes $n^{-1}(\sigma_\xi^2 + \sigma_e^2)$ which can be reduced to $n^{-1} \sigma_e^2$ if the correct form of the regression line can be fitted. In case this correct model is used to construct the regression estimator, one is required to know additional population data. For instance, for a quadratic regression, one has to know $\bar{M}_{(p)}$ and $\Sigma_{i=1}^{N}(M_i - \bar{M}_{(p)})^2$. There is no difficulty in constructing the quadratic regression estimator if each population cluster size is known. Regarding the variance estimation, the sample residual mean square error

s_d^2 is a biased estimate of $\sigma_y^2(1 - \rho^2)$ when the population regression is nonlinear, but the bias is still of order n^{-1} .

Konijn (1962) proposed the model in which the M_0 individuals are treated as a proportional stratified sample from an infinite superpopulation of individuals. The regression model is

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + e_{ij} , \quad j = 1,2,\ldots, M_i ; \quad i = 1,2,\ldots, N ,$$

$$E(e_{ij} | X_{ij} = X_0) = 0 ,$$

$$E(e_{ij}^2 | X_{ij} = X_0) = \sigma_i^2 .$$

First, the n clusters are assumed to be selected by some known design $(p_1,\ldots, p_N)$, where p_i is the probability of selecting a cluster i . Let p_{ij} be the joint probability that cluster i and cluster j are sampled. After the i-th cluster sample is selected, a simple random sample S_i of m_i (> 1) subunits is taken independent of the first-stage sampling procedure. The objective is to estimate

$$\alpha = \left(\sum_{i=1}^{N} M_i \right)^{-1} \left(\sum_{i=1}^{N} M_i \alpha_i \right) ,$$

$$\beta = \left(\sum_{i=1}^{N} M_i \right)^{-1} \left(\sum_{i=1}^{N} M_i \beta_i \right) ,$$

and the variances of the estimators. Under the known design $(p_1,\ldots, p_n)$, Konijn proposed

$$\hat{\alpha}_i = \bar{Y}_{i.} - \hat{\beta}_i \bar{X}_{i.} ,$$

$$\hat{\beta}_i = [\sum_{j=1}^{m_i} (X_{ij} - \bar{X}_{i.})^2]^{-1} [\sum_{j=1}^{m_i} (X_{ij} - \bar{X}_{i.}) Y_{ij}] ,$$

$$\hat{\alpha} = \sum_{i=1}^{n} k_i \hat{\alpha}_i ,$$

$$\hat{\beta} = \sum_{i=1}^{n} k_i \hat{\beta}_i ,$$

where $k_i = (p_i M_0)^{-1} M_i$. The variance of $\hat{\beta}$ is given by

$$V(\hat{\beta}) = \sum_{i=1}^{N} p_i \sigma_i^2 k_i^2 F_i^2 + \sum_{i=1}^{N} p_i (1 - p_i) \beta_i^2 k_i^2$$

$$+ \sum_{i \neq j}^{N} \sum (p_{ij} - p_i p_j) \beta_i \beta_j k_i k_j ,$$

where

$$F_i = E_i \left[\frac{1}{\sum_{j=1}^{m_i} (X_{ij} - \bar{X}_{i.})^2} \right],$$

E_i = conditional mean given cluster i .

Therefore, when the first-stage sampling is simple random sampling, i.e., $p_i = \frac{n}{N}$ and $p_{ij} = \frac{n}{N}\frac{n-1}{N-1}$, the variance of $\hat{\beta}$ becomes

$$V(\hat{\beta}) = \frac{n}{N}\left[\sum_{i=1}^{N} \sigma_i^2 k_i^2 F_i^2 + \frac{N-n}{N-1} \sum_{i=1}^{N} (\beta_i k_i - N^{-1} \sum_{j=1}^{N} \beta_j k_j)^2\right] .$$

One can see that the first term in $V(\hat{\beta})$ is associated with the variability of $\hat{\beta}_i$ as an estimator of β_i , whereas the second term is due to the sampling variability of the β_i's .

Konijn also constructed an unbiased estimator of $V(\hat{\beta})$. For design $(p_1, \ldots, p_N)$, the variance estimator is given by

$$\hat{V}(\hat{\beta}) = \sum_{i=1}^{n}\left[\frac{p_i \hat{\sigma}_i^2 k_i^2}{\sum_{j=1}^{m_i} (X_{ij} - \bar{X}_{i.})^2}\right] + \hat{\beta}^2 - \sum_{i=1}^{n} p_i \hat{\beta}_j^2 k_i^2 - \sum_{i\neq j}^{n}\sum \frac{p_i p_j \hat{\beta}_i \hat{\beta}_j k_i k_j}{p_{ij}} ,$$

where

$$\hat{\sigma}_i^2 = (m_i - 2)^{-1} \sum_{j=1}^{m_i} (Y_{ij} - \hat{\alpha}_i - \hat{\beta}_i X_{ij})^2 .$$

For the case where $p_i = \frac{n}{N}$ and $p_{ij} = \frac{n(n-1)}{N(N-1)}$, this variance estimator reduces to

$$\hat{V}(\hat{\beta}) = \frac{n}{N}\left[\sum_{i=1}^{n} \frac{\hat{\sigma}_i^2 k_i^2}{m_i \sum_{j=1}^{} (X_{ij} - \bar{X}_{i.})^2} + \frac{N-n}{N-1} \sum_{i=1}^{n} (\hat{\beta}_i k_i - \frac{\hat{\beta}}{n})^2\right] .$$

Alternatively, Konijn dealt with the situation under which clusters are selected with replacement. Denote by p_i^{ℓ} the probability of including cluster i as the ℓ-th selected cluster; let $\bar{p}_i = n^{-1}(p_i^1 + \ldots + p_i^n)$. This $\bar{p}_i$ may be interpreted as the average probability of selecting cluster i. Konijn then proposed

$$\hat{\beta} = \sum_{i=1}^{n} k_i' \hat{\beta}_i, \qquad \hat{\alpha} = \sum_{i=1}^{n} k_i' \hat{\alpha}_i ,$$

where $k_i' = \frac{M_i}{n \bar{p}_i M_0}$. Note that here the summation is over the sequence of selected clusters, so that several of the $\hat{\beta}_i$ (or $\hat{\alpha}_i$) may be identical. The variance of $\hat{\beta}$ is given by

$$V(\hat{\beta}) = n[\sum_{i=1}^{N} (\bar{p}_i + (n-1)\bar{p}_{ii})\sigma_i^2 k_i'^2 F_i^2 + \sum_{i\neq 1}^{N} \bar{p}_i \beta_i^2 k_i'^2]$$

$$+ (n-1) \sum_{i\neq j}^{N} \sum \bar{p}_{ij} \beta_i \beta_j k_i' k_j' - \beta^2 ,$$

where $\bar{p}_{ij}$ is the average joint probability of selecting cluster i and cluster j. An unbiased estimate of $\hat{V}(\hat{\beta})$ is

$$\hat{V}(\hat{\beta}) = \sum_{i=1}^{n} \frac{n\, p_i \hat{\sigma}_i^2\, k_i'^2}{\sum_{j=1}^{n} (X_{ij} - \bar{X}_{i.})^2} + \hat{\beta}^2 - \frac{n}{n-1} \sum_{i \neq j}^{n} \sum \frac{\bar{p}_i \bar{p}_j \hat{\beta}_i \hat{\beta}_j k_i' k_j'}{\bar{p}_{ij}} .$$

Fuller and Battese (1973) presented the use of transformations for the estimation problem in a linear model with nested-error structure. The one-fold nested-error model is expressed as

$$Y_{ij} = \underset{\sim}{x}_{ij}\underset{\sim}{\beta} + u_{ij} , \quad j = 1,2,\ldots, m_i , \quad i = 1,2,\ldots, n ,$$

$$u_{ij} = v_i + e_{ij} , \qquad (2.B.1)$$

where Y_{ij} denotes the Y-value of the j-th element in cluster i ; $\underset{\sim}{x}_{ij}$ denotes $1 \times p$ vector of auxiliary measurements for the j-th element in cluster i ; v_i is the i-th cluster effect; and e_{ij} is a random error. It is assumed that v_i and e_{ij} are independently distributed with zero means and variances σ_v^2 and σ_e^2 , respectively, where $\sigma_v^2 > 0$ and $\sigma_e^2 > 0$. The linear model (2.B.1) can be rewritten as

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{u} ,$$

$$E(\underset{\sim}{u}\,\underset{\sim}{u}') = \underset{\sim}{V} = \text{Block diag}[\underset{\sim}{V}_1, \underset{\sim}{V}_2, \ldots, \underset{\sim}{V}_n] , \qquad (2.B.2)$$

where

$$\underset{\sim}{Y} = (\underset{\sim}{Y}_1', \ldots, \underset{\sim}{Y}_n')' ,$$

$$\underset{\sim}{Y}_i' = (Y_{i1}, \ldots, Y_{im_i}) ,$$

$\underset{\sim}{X}$ and $\underset{\sim}{u}$ are constructed in a similar manner, and $\underset{\sim}{V}_i = \sigma_e^2 I_{m_i} + \sigma_v^2 J_{m_i}$ with I_{m_i} denoting the identity matrix of order m_i and J_{m_i} denoting the $(m_i \times m_i)$ matrix with all elements equal to one. The transformation that makes the transformed errors uncorrelated with variances equal to one is given by

$$\underset{\sim}{T} = \begin{bmatrix} \underset{\sim}{T}_1 & 0 & \ldots & 0 \\ 0 & \underset{\sim}{T}_2 & & 0 \\ 0 & 0 & \ldots & \underset{\sim}{T}_n \end{bmatrix} ,$$

where $\underset{\sim}{T}_i = \sigma_e^{-1} I_{m_i} - m_i^{-1}\{\sigma_e^{-1} - (\sigma_e^2 + m_i\sigma_v^2)^{-1/2}\}J_{m_i}$. In practice, σ_e^2 and σ_v^2 are not known. In such a case, σ_e^2 and σ_v^2 are estimated by the "fitting-of-constants" method.

The resulting estimated generalized least-squares estimator is

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\hat{\underset{\sim}{V}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\hat{\underset{\sim}{V}}^{-1}Y ,$$

where $\hat{V}$ is obtained by substituting $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$ for σ_e^2 and σ_v^2 , respectively. The unbiasedness of $\hat{\beta}$ is demonstrated under the

assumptions that the errors u_{ij} are symmetrically distributed with fourth moments and $E[(\hat{\sigma}_e^2)^{-1}]$ exists. To investigate desirable large-sample properties, Fuller and Battese imposed the following conditions:

(1) u_{ij} have finite fourth moments,

(2) n^{-1} and $[\sum_{i=1}^{n} (m_i - 1)]^{-1}$ are both of order $[\sum_{i=1}^{n} m_i]^{-\delta}$, where $\delta > 0$,

(3) $\underset{\sim}{X}'\underset{\sim}{X}$ is nonsingular for all n,

(4) $\lim_{n\to\infty} n^{-1} \underset{\sim}{X}'\underset{\sim}{X}$ and $\lim_{n\to\infty} n^{-1} \underset{\sim}{X}'\underset{\sim}{V}^{-1} \underset{\sim}{X}$ exist and are positive definite.

It follows that $\hat{\underset{\sim}{\beta}}$ has the same asymptotic distribution as the estimator $\tilde{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{V}^{-1} \underset{\sim}{X})^{-1} \underset{\sim}{X}'\underset{\sim}{V}^{-1} \underset{\sim}{Y}$ under these conditions. The computational algorithm has been programmed in the computer package SUPER CARP developed by Hidiroglou et al. (1980).

In recent decades, much attention has been given to the effects of sampling design on the complex statistics such as differences between domain means, correlation coefficients, and regression coefficients. Frankel (1971), Kish and Frankel (1974) have given empirical results for the estimation of these complex statistics in two-stage sampling from a fixed finite population. No population model was assumed. In their studies, the term "first-order statistics" was used to denote sample estimates of parameters of the population distribution (e.g., means,

correlation coefficients, regression coefficients), and "second-order statistics" was used to denote estimates of the sampling variability of the first-order estimates.

Their empirical studies show that the relative biases for all five types of estimators (i.e., regression coefficient, partial correlation coefficient, simple correlation coefficient, multiple correlation coefficient, and ratio means) are small and decrease as the sample size increases. The multiple correlation coefficients often have very large relative bias for small samples.

Besides small biases, the approaches to normality of complex statistics from two-stage samples are good even in moderate sample sizes. The proportion of times that the ratio of the first-order estimate minus its expected value to its estimated standard error falls within symmetric intervals about the origin is about equal to the proportion predicted by a standard normal. The rate of the approach to population values is affected by positive correlations among selected elements.

Measures of variation of first-order estimates are affected by correlations between elements which have been induced by the sampling designs. To describe complex sampling designs, Kish (1965) defined a quantity, the design effect (Deff), as the ratio of the variance of the estimator obtained from a complex sampling design to the variance obtained from a simple random sample of the same number of units. In terms of design effects, Kish and Frankel (1974) made conjectures on the

variances of complex statistics from complex samples. Their conjectures are summarized as follows:

(i) Standard errors computed based on simple random assumptions tend to underestimate the standard errors of complex statistics from complex samples;

(ii) The design effects for second order statistics tend to be less than those for means of the same variables. Also, the higher the design effects for the latter, the higher the design effects for the former tend to be.

Three basic methods of producing second-order estimates were studied: the Taylor expansion method (TAYLOR), balanced repeated replication method (BRR), and jackknife repeated replication method (JRR). Criteria used for comparisons among these three methods were the relative bias and the proportion of times that the t-ratio of the first-order estimate falls within a symmetric interval about the expected value. Kish and Frankel (1974) found that with the exception of estimators for multiple correlation coefficients, all three methods give small relative biases for the estimators of the mean square errors of the complex statistics. The t-ratio values agree well with the Student's expected probabilities. As expected, the relative biases and t-ratio values improve with increasing sample size. The empirical study also suggests that strong positive correlations between the selected elements result in large numerators and denominators of the t

ratios. Finally, it was concluded from the empirical studies that none of the three methods dominates the others consistently. Therefore, the choice among these methods will depend on the statistics used, simplicity, and costs.

Campbell (1977) studied the design effect of the ordinary least squares estimator $\hat{\beta}_{OLS}$ of β in the infinite population model (2.B.1) considered by Fuller and Battese (1973). The X-variable is allowed to be a random variable. The random error $\underset{\sim}{u}$ in model (2.B.2) is unobserved with $E(\underset{\sim}{u}|\underset{\sim}{X}) = 0$ and $var(\underset{\sim}{u}|\underset{\sim}{X}) = \sigma^2 \underset{\sim}{V}$, defined in model (2.B.2). In the simple linear regression case (i.e., $p = 1$), Campbell made the following points:

(i) $Deff(\hat{\beta}_{OLS}|\underset{\sim}{X}) = 1 + (Deff(\underset{\sim}{1}'\underset{\sim}{X}) - 1)\rho$, where $Deff(\underset{\sim}{1}'\underset{\sim}{X})$ is the design effect for estimating the total of $\underset{\sim}{X}$ given by

$$Deff(\underset{\sim}{1}'\underset{\sim}{X}) = 1 + \left(\frac{Var(m_i)}{\bar{m}} + \bar{m} - 1 \right) \rho_X ,$$

$\bar{m}$ being the average sampled cluster size, ρ_X being the intracluster correlation for X, and ρ being the intracluster correlation of the residuals around the regression line. When sampled cluster sizes are equal, $Deff(\hat{\beta}_{OLS}|\underset{\sim}{X}) = 1 + (\bar{m} - 1)\rho_X\rho$.

(ii) If $\rho_X > 0$ and $\rho > 0$, then

$$Deff(\hat{\beta}_{OLS}|\underset{\sim}{X}) > 1 ,$$

$$\text{Deff}(\hat{\beta}_{OLS}|\underset{\sim}{X}) \leq \text{Deff}(\underset{\sim}{1}'\underset{\sim}{X}) ,$$

and

$$\text{Deff}(\hat{\beta}_{OLS}|\underset{\sim}{X}) \leq \text{Deff}(\underset{\sim}{1}'\underset{\sim}{Y}) .$$

The points in (ii) give theoretical evidence to support Kish and Frankel's (1974) conjectures. The presence of an additional independent variable Z in the model will naturally create complexities due to the correlation between X and Z. The conditional design effect for the OLS estimator $\hat{\beta}_X$ of the coefficient corresponding to X is then given by

$$\text{Deff}(\hat{\beta}_X|\underset{\sim}{X}, \underset{\sim}{Z})$$

$$= 1 + \left[\frac{\text{Deff}(\bar{\bar{x}}) - 2\gamma\, \rho_{ZX}\sqrt{\text{Deff}(\bar{\bar{x}})\text{Deff}(\bar{\bar{z}})} + \gamma^2\, \text{Deff}(\bar{\bar{z}})}{1 - \gamma^2} - 1 \right] \rho ,$$

where $\bar{x}$ and $\bar{z}$ are the sample means of X and Z, respectively, ρ_{ZX} is the correlation coefficient between cluster totals for Z and X, and γ is the correlation coefficient between elements for Z and X.

The efficiency of the ordinary least squares estimator relative to the generalized least squares estimator was also studied. The ordinary least squares estimator appears reasonably efficient for small values of ρ and ρ_X when the average sampled cluster size $\bar{m}$ is no greater than 50.

The effect of the intracluster correlation on the ordinary least squares estimation procedure was also examined by Holt and Scott (1981). The simple linear regression model for the case where $m_i = m$ for all clusters can be written in the form

$$Y_{ij} = \alpha + \beta(X_{ij} - \bar{X}_{..}) + u_{ij} \ , \quad j = 1,2,\ldots, m \ , \quad i = 1,2,\ldots, n \ ,$$

where

$$\bar{X}_{..} = n^{-1} m^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij} = m^{-1} \sum_{i=1}^{n} \bar{X}_{i.}$$

and $\bar{X}_{i.}$ is the i-th cluster mean of X. The design effect of the OLS estimator $(\hat{\alpha}_{OLS}, \hat{\beta}_{OLS})'$ is given by

$$D = \begin{pmatrix} 1 + (m - 1)\rho & 0 \\ 0 & 1 + (m - 1)\hat{\rho}_X\rho \end{pmatrix} ,$$

where

$$B_X = m \sum_{i=1}^{n} (\bar{X}_{i.} - \bar{X}_{..})^2 ,$$

$$T_X = \sum_{i=1}^{c} \sum_{j=1}^{m} (X_{ij} - \bar{X}_{..})^2 ,$$

$$\hat{\rho}_X = \left(\frac{m B_X}{T_X} - 1 \right) (m - 1)^{-1} ,$$

the sample intracluster correlation of X. On the other hand, the model which permits different within and between cluster regressions may be expressed in the following form

$$Y_{ij} = \alpha + \beta_w (X_{ij} - \bar{X}_{i.}) + \beta_b (\bar{X}_{i.} - \bar{X}_{..}) + u_{ij} .$$

Under this model, the design effect of the OLS estimator $(\hat{\alpha}, \hat{\beta}_w, \hat{\beta}_b)'$ is given by

$$D = \begin{bmatrix} 1 + (m - 1)\rho & 0 & 0 \\ 0 & 1 - \rho & 0 \\ 0 & 0 & 1 + (m - 1)\rho \end{bmatrix} .$$

In unbalanced cases where the cluster sample sizes are not all equal, the diagonal form of both D matrices is destroyed. From these two diagonal matrices, one can easily see that the clustering effect on $V(\hat{\beta}_{OLS})$ will fall between the clustering effect on

$V(\hat{\beta}_w)$ and on $V(\hat{\beta}_b)$. Furthermore, compared to $V(\hat{\beta}_{OLS})$, the clustering effect on $V(\hat{\beta}_b)$ is an increase, whereas that on $V(\hat{\beta}_w)$ is a decrease. The clustering has a much larger effect on $V(\hat{\alpha}_{OLS})$ than on $V(\hat{\beta}_{OLS})$, which is consistent with the results of Campbell (1977). For the general linear model, $Y = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{u}$, with p independent variables, $\{X_1, \ldots, X_p\}$, the covariance matrix of the ordinary least squares estimators $\hat{\underset{\sim}{\beta}}_{OLS}$ is given by

$$V(\hat{\underset{\sim}{\beta}}_{OLS}) = \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1} D_3 ,$$

where $D_3 = I + (M^* - I)\rho$ and

$$M^* = \{ \sum_{i=1}^{n} m_i \underset{\sim}{X}'_{B_i} \underset{\sim}{X}_{B_i} \}(\underset{\sim}{X}'\underset{\sim}{X})^{-1} .$$

(Here $\underset{\sim}{X}_{B_i}$ represents the $m_i \times p$ matrix with every element in the ℓ-th column equal to the average value of X_ℓ over the i-th cluster.)

The efficiency of $\hat{\underset{\sim}{\beta}}_{OLS}$ relative to the generalized least squares (GLS) estimator $\hat{\underset{\sim}{\beta}}_G$ was investigated by Scott and Holt (1982). Let $e(\underset{\sim}{C})$ denote the ratio of $V(\underset{\sim}{C}'\hat{\underset{\sim}{\beta}}_G)$ to $V(\underset{\sim}{C}'\hat{\underset{\sim}{\beta}}_{OLS})$ for an arbitrary constant vector $\underset{\sim}{C}$. Then, by assuming that $\rho > 0$, the upper bound on the loss of efficiency, $1 - e(\underset{\sim}{C})$, is given by

$$L_{max} = \left(1 + \frac{4(1 - \rho)[1 + (m - 1)\rho]}{m^2 \rho^2} \right)^{-1} .$$

When the cluster sample sizes are not all equal, the term m in this bound is replaced by m_{max}, the maximum cluster sample size. As Scott and Holt noted, if the value of ρ is reasonably small, the corresponding loss in efficiency is not large. In their experience, a large value for ρ is often a warning that an important explanatory variable has been left out of the model.

The usual estimator of σ^2 tends to slightly underestimate σ^2, but the effect will be negligible if the sample size is reasonably large. The main failure of confidence intervals and test procedures based on OLS results is due to using $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ in place of $(\underset{\sim}{X}'\underset{\sim}{X})^{-1} D$. Holt and Scott (1981) presented an example which showed that in testing $H_0: \underset{\sim}{C}'\underset{\sim}{\beta} = d_0$ at the nominal significance level of 5 percent, the coverage probability of the usual t test may be 74 percent of the nominal coverage rate for $m = 11$ and $\rho = 0.10$.

DeMets and Halperin (1977) dealt with the problem of estimation of regression coefficients under the situation where a finite population of size N is a simple random sample from a superpopulation and a sample of size n is drawn from these N units. A variable Z, referred to as a design variable, is known at the design stage for each member of the finite population. This design variable Z is employed to

determine the sample design $p(s)$ so that the sample is selected using the design $p(s|Z)$. For example, Z serves as a grouping variable for stratified or cluster sampling. After sampling, observations are made on the dependent variable Y and on the independent variable X. It was assumed that (Y, X, Z) in the infinite population is distributed as a multivariate normal vector with mean vector (μ_Y, μ_X, μ_Z) and positive definite covariance matrix Σ, where

$$\Sigma = \begin{bmatrix} \sigma_Y^2 & \rho_{YX}\sigma_Y\sigma_X & \rho_{YZ}\sigma_Y\sigma_Z \\ \rho_{YX}\sigma_Y\sigma_X & \sigma_X^2 & \rho_{XZ}\sigma_X\sigma_Z \\ \rho_{YZ}\sigma_Y\sigma_Z & \rho_{XZ}\sigma_X\sigma_Z & \sigma_Z^2 \end{bmatrix}.$$

The main parameter of interest is the infinite population regression coefficient $\beta = \rho_{YX}\sigma_Y/\sigma_X$. The usual ordinary least squares estimator (OLS) of β is $\hat{\beta}_{OLS} = s_{XY}/s_X^2$, which is asymptotically biased unless the sample variance of Z approaches σ_Z^2 as the sample size n increases. The maximum likelihood estimator (MLE) under the trinormal distribution of (Y, X, Z) was proposed and shown to be an asymptotically unbiased estimator. The empirical results reveal that, in general, the variance of MLE is less than the variance of OLS for ρ_{XY} smaller than 0.8 and slightly greater otherwise. The mean square error of MLE is less than that of OLS.

Nathan and Holt (1980) give a set of linear model assumptions which are weaker than the trinormality assumption. These assumptions can be summarized as follows:

(i) The conditional expectations of Y and X given Z are linear in Z ;

(ii) The conditional covariance matrix of Y and X given Z does not depend on Z ;

(iii) Given all the finite population values of the design variable Z , the W's are conditionally independent for different units, where $W = (Y, X)'$.

Under these assumptions, the usual OLS estimator $\hat{\beta}_{OLS}$ of β is asymptotically biased conditional on the sample and all the finite population values of Z . The unconditional bias of $\hat{\beta}_{OLS}$, under the situation when the variance of sample variance s_Z^2 is $O(n^{-1})$, cannot be ignored for large n unless: (i) the design variable Z is noninformative about X (i.e., $\rho_{XZ} = 0$) ; (ii) Z provides no further information on Y than that provided by X (i.e., the partial correlation $\rho_{YZ.X}$ given X is zero); or (iii) $E(s_Z^2) = \sigma_Z^2$,and this holds exactly for simple random sampling.

The maximum likelihood estimator $\hat{\beta}_{MLE}$ adopted by DeMets and Halperin (1977) has also been shown to be asymptotically unconditionally unbiased under these weaker assumptions. Under conditions where $\hat{\beta}_{OLS}$ is asymptotically unbiased, neither $\hat{\beta}_{OLS}$ nor $\hat{\beta}_{MLE}$ has

uniformly smaller variance; however, if $E(s_Z^2) = \sigma_Z^2$, then $\hat{\beta}_{MLE}$ has smaller variance than $\hat{\beta}_{OLS}$.

Two weighted estimators were also considered by Nathan and Holt (1980). These estimators are based on weighted sample means, variances, and covariances where the weights are the inverses of the sample inclusion probabilities. Each inclusion probability is defined to be $\pi_\alpha = p(\alpha \in S|Z)$ and is assumed to be greater than zero for each population unit. The two weighted estimators, $\hat{\beta}^*_{OLS}$ and $\hat{\beta}^*_{MLE}$, are obtained by substituting the weighted statistics for their unweighted counterparts. Properties of these two weighted estimators are summarized as follows:

(i) To the term of $O(n^{-1})$, the unconditional variances of these two weighted estimators are greater than that of $\hat{\beta}_{MLE}$ for PPS sampling with replacement.

(ii) To the term of $O(n^{-1})$, $\hat{\beta}^*_{OLS}$ and $\hat{\beta}^*_{MLE}$ are design unbiased for the finite population regression coefficient. That is, given $(Y_1, X_1, Z_1),\ldots, (Y_N, X_N, Z_N)$, the expectations of these two estimators taken over all possible samples will be equal to the finite population regression coefficient. This property of design unbiasedness yields asymptotic unconditional unbiasedness for $\hat{\beta}^*_{OLS}$ and $\hat{\beta}^*_{MLE}$.

(iii) $\hat{\beta}^*_{OLS}$ and $\hat{\beta}^*_{MLE}$ are model free and, hence, may be more robust to departures from the model than the unweighted estimators.

The design variable Z can also be used for stratification. For the vector (Y, X, Z) define finite population means $\bar{Y}$, $\bar{X}$, $\bar{Z}$; finite population variances $\hat{\sigma}^2_Y$, $\hat{\sigma}^2_X$, $\hat{\sigma}^2_Z$; sample means $\bar{y}$, $\bar{x}$, $\bar{z}$; and sample variances and covariances s^2_x, s^2_y, s^2_z, s_{xy}, s_{xz}, s_{yz} . The conditional variance of Y given X and Z is estimated as follows:

$$\hat{\sigma}^2_{y.xz} = s^2_y - (s^2_{yz}s^2_x + s^2_{yx}s^2_z - 2s_{yx}s_{yz}s_{xz})(s^2_x s^2_z - s^2_{xz})^{-1} .$$

For each stratum h , let $W_h = N_h/n_h$, let (Y_{hi}, X_{hi}, Z_{hi}) refer to the value of (Y, X, Z) for the i-th individual in the h-th stratum, and let $\sum_{i\varepsilon s}$ denote the sum over the individuals in the sample. Let

$$U_{hi} = W_h(X_{hi} - \hat{\bar{X}})\{Y_{hi} - \hat{\bar{Y}} - \hat{\beta}^*(X_{hi} - \hat{\bar{X}})\}(\sum_h \sum_{i\varepsilon s} W_h(X_{hi} - \hat{\bar{X}})^2)^{-1},$$

$$N\hat{\bar{Y}} = \sum_h \sum_{i\varepsilon s} W_h Y_{hi} ,$$

$$N\hat{\bar{X}} = \sum_h \sum_{i\varepsilon s} W_h X_{hi} ,$$

where n_h and N_h are sample size and population size for the i-th stratum, with $n = \sum_h n_h$, and $N = \sum_h N_h$.

Holt et al. (1980) conducted a series of empirical studies for the case where the design variable Z is used to construct strata. Three procedures for estimating the regression coefficient were compared. Table 2.1 lists the estimators and their associated variance estimates. Two real data sets were used to obtain realistic values of population parameters, such as variances and covariances, for the computer simulation. A finite population of 10,000 numbers was generated for the design variable Z using a random normal generator. From this population, various sample designs were used to select samples of size 1,000: (i) simple random sampling, (ii) proportionate stratification allocation, (iii) increasing allocation, (iv) U-shaped allocation. The design variable Z stratified the population into five equal strata of size 2,000 each. Sample allocations for (ii), (iii), and (iv) were as follows:

(ii) proportionate stratified allocation (200, 200, 200, 200 200);

(iii a) increasing allocation (50, 150, 200, 250, 300);

(iii b) increasing allocation (50, 50, 100, 300, 500);

(iv a) U-shaped allocation (300, 150, 100, 150, 300);

(iv b) U-shaped allocation (450, 49, 2, 49, 450).

Note that the U-shaped allocation designs tend to select two extremes of the Z values, while the increasing allocation designs tend to select

Table 2.1. Three procedures for estimating the regression coefficient

Procedures	Estimator	Estimated Variance
OLS	$\hat{\beta}_{OLS} = s_{yx}/s_x^2$	$(s_y^2 - \hat{\beta}_{OLS} s_{xy})(n\, s_x^2)^{-1}$
MLE	$\hat{\beta}_{MLE} = \left\{ s_{yx} + \frac{s_{yz} s_{xz}}{s_z^2} \left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1 \right) \right\} \times \left\{ s_x^2 + \frac{s_{xz}^2}{s_z^2} \left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1 \right) \right\}^{-1}$	$\hat{\sigma}_{y.xz}^2 \left\{ s_x^2 + \frac{s_{xz}^2}{s_z^2} \left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1 \right) \right\}^2 + 2 \frac{s_{xz}^2}{s_z^2} \left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1 \right) \times n^{-1} \left(s_x^2 + \frac{s_{xz}^2}{s_q^2} \left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1 \right) \right)^{-2}$
p-weighted	$\hat{\beta}^* = \left[\sum_h \sum_{i \in s} W_h (Y_{hi} - \hat{\bar{Y}})(X_{hi} - \hat{\bar{X}}) \right] \times \left[\sum_h \sum_{i \in s} W_h (X_{hi} - \hat{\bar{X}})^2 \right]^{-1}$	$\sum_h \frac{n_h^2}{n_h - 1} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sum_{i \in s} (U_{hi} - \bar{U}_h)^2$

the largest values of the design variable Z. For the selected Z values, a random normal generator generated the values for the independent variable X conditional on Z, and finally generated the values for the dependent variable Y conditional on the selected X and Z values. For each sample of 1,000 Z values, 10 replications of X were generated and for each of these replications, 100 replications of Y.

Holt et al. (1980) examined the bias and standard error for each of the three estimators. Results of the computer simulation can be summarized as follows:

(i) $\hat{\beta}_{OLS}$ performs satisfactorily with equal probability sampling, but $\hat{\beta}_{OLS}$ is biased for unequal probability samples with the bias sometimes more than 10 percent of the estimated value and much larger than the standard error;

(ii) $\hat{\beta}_{MLE}$ performs well in general. The U-shaped allocations demonstrate the efficiency for $\hat{\beta}_{MLE}$; whereas, $\hat{\beta}_{OLS}$ appears most subject to bias;

(iii) The p-weighted estimator $\hat{\beta}^*$ performs well in terms of bias but is less stable in terms of standard error. In some situations, $\hat{\beta}^*$ has variance which is ten times larger than that of $\hat{\beta}_{MLE}$. Even for the U-shaped allocations, $\hat{\beta}^*$ does not appear to have smaller standard error.

Holt et al. (1980) also examined average frequencies for interval estimates of β. The interval estimator of β is usually given by

$$\hat{\beta} \pm z_{\alpha/2}[\hat{V}(\hat{\beta})]^{1/2},$$

where $\hat{\beta}$ is an estimator of β with the variance estimator $\hat{V}(\hat{\beta})$ and $z_{\alpha/2}$ is the upper $100\ \alpha/2$ percent point of a standard normal distribution. The distributions of coverage frequency (the frequency that the confidence intervals of the form $\hat{\beta} \pm z_{\alpha/2}(v(\hat{\beta}))^{1/2}$ cover the true value β) were compared to the expected frequency for each estimator of β and each design. Holt et al. (1980) noted the following:

(i) For unequal probability designs, the coverage properties for $\hat{\beta}_{OLS}$ are poor as expected;

(ii) Except for the most extreme allocation (iv b), the coverge properties of $\hat{\beta}_{MLE}$ and $\hat{\beta}^*$ are acceptable, but intervals based on $\hat{\beta}_{OLS}$ and the OLS estimator of variance are acceptable only for equal probability designs;

(iii) For the most extreme allocation (iv b), the coverage properties of p-weighted estimators are poor with high frequencies in the tails of the distribution. Holt et al. (1980) also pointed out that the p-weighted estimator will place the most weight on the central stratum and the

least weight on the extreme strata in the extreme design. If the design variable Z is positively correlated with the independent variable X, the weighting process for the $\hat{\beta}^*$ will conflict with the design itself in which the most extreme strata should receive most weight. Therefore, the p-weighted procedure cannot be used to improve efficiency under the heavily unequal probability designs.

Fuller (1975) investigated the asymptotic properties of the estimator of the finite population regression coefficient. The finite population $\{(Y_t, X_{1t}, X_{2t}, \ldots, X_{pt}): t = 1,2,\ldots, N\}$ is assumed to be a random sample from a multivariate infinite population with finite fourth moments and a positive definite covariance matrix. The vector of the finite population regression coefficients is defined by

$$\underset{\sim}{B} = \left(\sum_{t=1}^{N} \underset{\sim}{X}_t' \underset{\sim}{X}_t \right)^{-1} \sum_{t=1}^{N} \underset{\sim}{X}_t' Y_t$$

and the infinite population vector of coefficients by

$$\underset{\sim}{\beta} = \left(E\{\underset{\sim}{X}_t' \underset{\sim}{X}_t\}\right)^{-1} E\{\underset{\sim}{X}_t' Y_t\} ,$$

where

$$\underset{\sim}{X}_t = (X_{1t}, \ldots, X_{pt})' , \quad t = 1,2,\ldots, N .$$

The sample estimator of $\underset{\sim}{\beta}$, based on a simple random sample of size n , is given by

$$\hat{\underset{\sim}{\beta}} = \left(\sum_{t=1}^{n} \underset{\sim}{X}_t' \underset{\sim}{X}_t \right)^{-1} \sum_{t=1}^{n} \underset{\sim}{X}_t' Y_t \ .$$

As n , $N \to \infty$,

$$n^{1/2} (\hat{\underset{\sim}{\beta}} - \underset{\sim}{B}) \xrightarrow{L} N(\underset{\sim}{0}, \ (1 - f) Q^{-1} G Q^{-1}) \ ,$$

and

$$n^{1/2} (\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}) \xrightarrow{L} N(\underset{\sim}{0}, \ Q^{-1} G Q^{-1}) \ ,$$

where f is the limit of the finite correction factor n/N , with

$$0 \leq f < 1 \ ,$$

$$Q = E(\underset{\sim}{X}_t' \underset{\sim}{X}_t) \ ,$$

and

$$G = E[\underset{\sim}{X}_t' \underset{\sim}{X}_t (Y_t - \underset{\sim}{X}_t' \underset{\sim}{\beta})^2] \ .$$

In the absence of the usual assumptions of the linear model, a consistent estimator of G is

$$\hat{G} = (n - p)^{-1} \sum_{i=1}^{n} \underset{\sim}{X}_t'(Y_t - \underset{\sim}{X}_t'\hat{\beta})^2 \underset{\sim}{X}_t \ .$$

In two-stage stratified sampling, a sample can usually be expressed as $\{(Y_{ijt}, X_{1ijt}, \ldots, X_{pijt}): t = 1,2,\ldots, m_{ij}; j = 1,2,\ldots, N_i; i = 1,2,\ldots, L\}$, where the finite population is divided into L strata of which the i-th stratum contains N_i primary units with M_{ij} elements within the j-th primary unit $(i = 1,2,\ldots, L; j = 1,2,\ldots, N_i)$. The usual estimator of the regression coefficient $\underset{\sim}{\beta}$ can be constructed by first weighting each observation within the j-th primary unit of the i-th stratum by

$$\omega_{ij} = \left(\frac{W_i}{n_i} \frac{M_{ij}}{m_{ij}} \right)^{1/2} ,$$

where

W_i = fraction of population in stratum i ,

n_i = number of primaries selected from stratum i

and

$$m_{ij} = \text{number of sample elements in primary } j \text{ of stratum } i .$$

For each i,j,t, let

$$\tilde{\underset{\sim}{X}}_{ijt} = (\omega_{ij}X_{1ijt}, \ldots, \omega_{ij}X_{pijt})' ,$$

let

$$\tilde{Y}_{ijt} = \omega_{ij}Y_{ijt} .$$

The resulting estimator is then given by

$$\hat{\underset{\sim}{\beta}}_W = [\sum_{i=1}^{L}\sum_{j=1}^{n_i}\sum_{t=1}^{m_{ij}} (\tilde{\underset{\sim}{X}}'_{ijt}\tilde{\underset{\sim}{X}}_{ijt})]^{-1}[\sum_{i=1}^{L}\sum_{j=1}^{n_i}\sum_{t=1}^{m_{ij}} (\tilde{\underset{\sim}{X}}'_{ijt}\tilde{Y}_{ijt})] .$$

The desirable asymptotic properties for $\hat{\underset{\sim}{\beta}}_W$ also follow by imposing sufficient conditions required to employ the Liapounov Central Limit Theorem. The estimator $\hat{\underset{\sim}{\beta}}_W$ and the estimator of the variance of $\hat{\underset{\sim}{\beta}}_W$ can be found in the SUPER CARP manual.

C. LANDSAT Crop Estimation

The Statistical Reporting Service (SRS) of the U.S. Department of Agriculture (USDA) is conducting research into the use of LANDSAT (land observatory satellite) data as auxiliary information to improve crop

estimation in agricultural surveys. One of the research areas is to improve estimates of crop acreages for multi-county areas, such as Crop Reporting Districts. Two sources of data are employed. The ground survey data taken in the annual agricultural survey, called the June Enumerative Survey (JES), serves as the primary survey variable for estimating crop acreages, and the satellite data serves as the auxiliary variable.

In the JES survey, the area-frame sampling design is utilized to create a sample which is a stratified area cluster sample. As Sigman, et al. (1978) described, two levels of stratification are generally used. The first-level strata are the individual states. Within each state, aerial photography is used to define the secondary strata, based on the percent of cultivated land. Table 2.C.1 gives an example of the stratum definitions in the state of Illinois. Within each stratum, the total area is divided into N_h area frame units called segments. Each segment is a well-defined land area about one-square mile in size. A simple random sample of n_h segments is drawn within each stratum.

During the ground survey visit, the acres devoted to each crop or land use are recorded for every field in each segment. In fact, broader information such as agricultural labor, grain storage on farms, livestock inventory, and so forth is also collected. Field boundary coordinates are related, through a computer process, to a map base so that very precise area measurements are available for individual fields. Through the maps the ground survey data set provides the

Table 2.C.1. Stratum numbers and definitions

Stratum		Substratum		
Description		Description		
10	intensive agriculture	11	75% + cultivated	
		12	50% - 75% cultivated	
50	nonintensive agriculture	20	15% - 49% cultivated	
		31, 32, 33	urban	noncultivated
		40	rangeland	noncultivated
		61	proposed water	noncultivated
		62	water	noncultivated

location of each field and of each segment to interface well with the satellite data.

LANDSAT satellite data consist of the set of measurements taken by the multispectral scanner system in the land observatory satellite. Values are obtained for an image area of approximately one acre called a pixel. This sensor system measures the amount of radiant energy reflected from the earth's surface in different wavelength bands of the spectrum. The LANDSAT II satellite has four bands: one green, one red, and two near-infrared bands.

The individual one acre areas, referred to as pixels, are arrayed along east-west rows within the 185 kilometer wide north-to-south pass of the LANDSAT satellite. Each pixel is recorded with four variables corresponding to the four bands. Several satellite passes are generally

required in order to cover an individual state. Different satellite passes have different image dates.

The satellite passes are split into scenes which are strips of LANDSAT data covering 185 × 185 square kilometer zones. Individual scenes are labeled by pass number and position (e.g., north, middle, south) in a pass. Adjacent scenes east-to-west overlap approximately one-third of the columns. The scene registration relates LANDSAT row/column coordinates to map based latitude/longitude by means of a linear regression equation. This registration process can determine whether or not the segment was correctly located and move the segment by row and column shifts if necessary. Ultimately, each segment is located with an accuracy of 1/2 pixel or better.

Each state is divided into nonoverlapping groups of contiguous counties wholly contained in a LANDSAT satellite pass. These county groups are called analysis districts and are determined by LANDSAT boundaries. Estimates are made for these analysis districts and the individual county estimates are developed from them.

Once the satellite data are accurately matched with the ground data, the LANDSAT data file provides the ground survey crop code and the signature (four band spectral readings) for each pixel. These multivariate measurements of four band radiometric readings are used to classify each pixel into a crop type by classification functions. This set of classification functions, referred to as the USDA pixel

classifier, corresponds one-to-one to a set of classification categories.

Swain and Davis (1978) reviewed the currently available quantitative approach to remote sensing. The commonly used technique is pattern recognition. The goal of this approach is to classify each elementary observation into one of a limited number of discrete classes. Swain and Davis (1978) present a model for a pattern-recognition system shown in Figure 2.C.1. The output of the sensor is a set of n measurements, each corresponding to one channel of the multispectral scanner with LANDSAT data. In LANDSAT II data, n is equal to four. In classification analysis, it is convenient to think of the n measurements as a point in n-dimensional space, often called the measurement space. The classifier assigns the measurement vector $\underset{\sim}{X} = (x_1, \ldots, x_n)'$ to one of a set of prespecified m classes, based on an appropriate classification rule. The measurement space is divided into decision regions, each corresponding to a specific class. A set

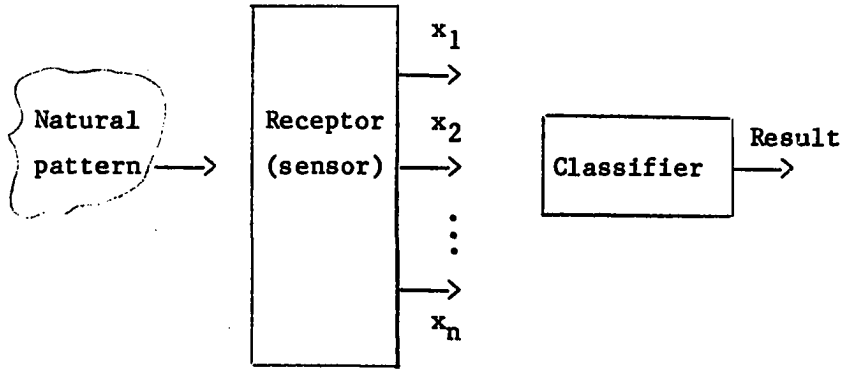

Figure 2.C.1. A model for a pattern recognition system

of m discriminant functions of $\underset{\sim}{X}$, denoted by $g_1(\underset{\sim}{X}),\ldots, g_m(\underset{\sim}{X})$, can be found so that whenever $\underset{\sim}{X}$ is a point in the i-th decision region, $g_i(\underset{\sim}{X})$ has a larger value than any other $g_j(\underset{\sim}{X})$, where $i,j \in \{1,2,\ldots, m\}$, $j \neq i$. In other words, the point $\underset{\sim}{X}$ is classified into the class with the largest $g_j(\underset{\sim}{X})$ value for $j \in \{1,2,\ldots, m\}$.

The discriminant functions are most often derived using the information obtained from a set of measurement vectors of known identity which are assumed to be representative of the classes of interest. This process of designing the classifier is often called training the classifier. The set of measurement vectors of known identity is called the training sample. Swain and Davis (1978) noted that when the measurement vectors of known identity are completely separable, deterministic classification functions can be found to correctly classify all of these measurement vectors. But when the classes of interest overlap in the measurement space, deterministic classification functions are generally not suitable. In such situations, statistical pattern recognition methods can be applied to develop classification functions.

In statistical pattern recognition procedures, probability functions associated with the classes are employed and must be estimated from a training sample. As in discriminant analysis, two methods are generally employed to estimate the unknown probability functions. Parametric methods are generally easier to implement, but require more

prior knowledge. Nonparametric methods are usually more powerful in the sense that they can more accurately estimate the probability functions, but this advantage is generally very expensive to achieve. Gleason et al. (1977), Craig et al. (1976), Sigman et al. (1978), and Amis et al. (1981) described the process of training the USDA pixel classifier. A sample of fields from each crop type is selected, and pixels inside the fields, called field-interior pixels, for a given cover type are extracted. The corresponding signatures are clustered in measurement space. Category likelihoods are computed by assuming that the signatures for a given crop category are distributed as a multivariate normal distribution. Thus, signature means and covariances and category prior probabilities from a training sample of labeled pixels are calculated. Once these parameters are estimated, all the pixels are classified.

Swain and Davis (1978) pointed out that the normal assumption usually provides a good trade-off between classification performance and cost. However, in cases where classes have distinctly multimodal probability distributions, a commonly practiced solution is to subdivide the class into a number of subclasses for which the probability function can be represented by a normal density function.

In discriminant analyses, the basic strategy a statistician follows is to minimize the average loss over the entire set of classifications to be performed. Such a strategy is often called Bayes optimal strategy. In mathematical formulation, let $p(\underset{\sim}{X}|\omega_i)$ be the probability

density function of measurement vector $\underset{\sim}{X}$, given that $\underset{\sim}{X}$ is from a pattern in class i; let $p(\omega_i)$ be the priori probability of class i. Let $C(j|i)$ be the cost of misclassifying an observation from class i as coming from class j, where $i,j \in \{i,2,\ldots, m\}$. Let $C(j|i) = 1$ if $i \neq j$ and $C(j|i) = 0$ otherwise. The Bayes optimal strategy leads us to the decision rule: assign $\underset{\sim}{X}$ to class i if and only if

$$p(\underset{\sim}{X}|\omega_i)p(\omega_i) \geq p(\underset{\sim}{X}|\omega_j)p(\omega_j) \quad \text{for all} \quad j = 1,2,\ldots, m .$$

This decision rule is called the maximum likelihood decision rule by Swain and Davis (1978). In the case of normal classes, the discriminant function for class i can be written as

$$g_i(\underset{\sim}{X}) = \ln p(\omega_i) - \tfrac{1}{2} \ln|\underset{\sim}{\Sigma}_i| - \tfrac{1}{2}(\underset{\sim}{X} - \underset{\sim}{\mu}_i)'\underset{\sim}{\Sigma}_i^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}_i) , \quad (2.C.1)$$

where

$$p(\underset{\sim}{X}|\omega_i) \sim N(\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma}_i) .$$

Another problem encountered in remote sensing is that there are inevitably a number of points which, in fact, do not belong to any of the classes under discussion. Swain and Davis (1978) introduced a well-known technique, called thresholding, to reject these points. In the

thresholding procedure, the threshold is specified by the user. If the probability values $p(\underset{\sim}{X}|\omega_i)$, $i = 1,2,\ldots, m$, are below that threshold, the data point is rejected. In particular, for the multivariate normal classes, a threshold level on $\underset{\sim}{X}$ can be converted to a threshold level on the quadratic function in $\underset{\sim}{X}$ in (2.C.1) by using the chi-square distribution.

The usual Bayes optimal discriminant rule tends to discriminate against classes with a low value for $p(\omega_i)$. If these "rare" classes are the main interest, Swain and Davis (1978) noted that the problem can be solved by using a loss function more complicated than the zero-one loss function. For example, one can use a loss function in which $C(j|i) = [p(\omega_i)]^{-1}$ if $i \neq j$ and $C(j|i) = 0$ otherwise.

The choice of a criterion to use for designing an effective classifier depends on the objective of the experiment. If the objective of the experiment is to "classify" the data points, the probability of error can be used as a criterion. In other words, the classifier is requested to minimize the overall probability of misclassification. In crop acreage estimation, however, the objective is to minimize the variance of resulting acreage estimates. The Bayes classification rule does not necessarily achieve this objective. For example, Gleason et al. (1977) investigated the effect on classifier performance of using "different prior probabilities" for the classification categories. Strictly speaking, there is only one correct set of prior probabilities for a given geographical region. Using "different prior probabilities"

actually means using different weighting factors for the likelihood functions in computing the class discriminant functions. In their study for Illinois corn acreage estimation, it was observed that equal prior probabilities yielded more precise crop acreage estimates compared to using probabilities defined as the ratio of the current year direct expanded acres to the total land area in the region. The current year direct expanded acres are computed as the sum of sample mean estimates for the total acres over all strata.

The regression estimation method has been used to improve the crop acreage estimation. USDA extracts information from LANDSAT by classifying individual signatures to probable crop type. Both ground data and classified LANDSAT data then can be utilized to estimate crop acreage by means of a regression estimator. Sigman et al. (1977, 1978) summarized the estimation procedure for a given state as follows: Let $h = 1,2,\ldots, L$ be L land-use strata. Within each stratum, the total area is divided into N_h segments from which a simple random sample of n_h segments is drawn. Let Y be total corn acres for a given state, and let y_{hj} be total corn acres in the j-th sampled segment in the h-th stratum. A regression estimator of the state corn total Y is

$$\hat{Y}_R = \sum_{h=1}^{L} N_h[\bar{y}_h + \hat{b}_h(\bar{X}_h - \bar{x}_h)] ,$$

where

$$\hat{b}_h = \left[\sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)^2\right]^{-1} \sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)(y_{hj} - \bar{y}_h) ,$$

X_{hi} = number of pixels classified as corn in the i-th segment of the h-th stratum ,

x_{hi} = number of pixels classified as corn in the i-th sampled segment in the h-th stratum ,

$$\bar{x}_h = n_h^{-1} \sum_{i=1}^{n_h} x_{hi} ,$$

$$\bar{X}_h = N_h^{-1} \sum_{i=1}^{N_h} X_{hi} ,$$

$$\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi} .$$

The estimated approximate variance for $\hat{Y}_R$ is given by

$$v(\hat{Y}_R) = \sum_{h=1}^{L} N_h^2 n_h^{-1} (1 - f_h) \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 (n_h - 2)^{-1} (1 - \hat{R}_h^2) ,$$

where

$$\hat{R}_h^2 = \left[\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2\right]^{-1} \left[\sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)^2\right]^{-1} \left[\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)\right]^2 ,$$

$$f_h = N_h^{-1} n_h \ .$$

Note that using only JES data, an estimate of total corn acres is

$$\hat{Y} = \sum_{h=1}^{L} N_h \bar{y}_h \ .$$

The estimated variance of the estimate $\hat{Y}$ is

$$v(\hat{Y}) = \sum_{h=1}^{L} N_h^2 n_h^{-1} (n_h - 1)^{-1}(1 - f_h) \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 \ .$$

Therefore, a substantially lower variance for $\hat{Y}_R$ is obtained if $\hat{R}_h^2$ is close to 1 for most strata. One can see that the auxiliary variables described above are

$$x_{hj} = \sum_k c(z_{hjk}) \quad \text{and} \quad X_{hj} = \sum_k c(Z_{hjk}) \ ,$$

where the variable $z_{hjk}(Z_{hjk})$ is the signature of the k-th pixel of the j-th sampled segment (j-th population segment) in the h-th stratum and the function $c(z)$ is 1 if the pixel with signature z is classified as corn and 0 otherwise. These auxiliary variables may not produce the estimate of Y with smallest possible variance. Hanuschak and Cardenas (1978) used a multiple regression estimator where the set of auxiliary variables also includes the classification results into cover types other than corn. Fuller (1977) suggested the use of the

posterior probability that a pixel with signature z is from corn. In this dissertation, we investigate the use of the posterior probability estimated by approximating it with the normal class-conditional probability density for each crop.

Stratum R-square values are often employed to study classifier performance. However, in small samples, the sample R-square values $\hat{R}_h^2$ can have a large positive bias as an estimate of the population squared coefficient of determination R_h^2. Sigman et al. (1978) and Gleason et al. (1975) found that in moderate size examples, e.g., $n_h = 84$, $\hat{R}_h^2$ is acceptable for estimating R_h^2. Hence, the additional labor involved in performing the jackknife calculations can be saved. Nevertheless, less biased estimates for R_h^2 can be obtained by using one of many methods used to estimate error rates in discriminant analysis. These methods were summarized by Toussaint (1974), e.g., jackknifing, sample partition, etc. Gleason et al. (1977) also studied three methods of estimating R_h^2. In the resubstitution method, all the segment data are used to both train and test the classifier. In the sample partition method, the classifier was trained on a 50 percent sample of segment fields, and then tested on all of the segment data. In jackknifing, the training set was 3/4 of the data, and the test data was the remaining 1/4. This allocation was repeated four times so that the union of the four test sets was the entire collection of segment data. Four separate estimates of classifier performance were obtained and averaged to yield the jackknife estimate. In general, the

resubstitution and sample partition methods are easy to perform, but they produce biased evaluations in small samples. The jackknife gives a less biased evaluation, but it involves more computational effort. The sensitivity of the classifier to the selection of the training data may be investigated by performing sample partition.

Mergerson (1981) studied the use of imagery from two different dates which cover the same land area in conjunction with JES ground data. It was shown that the use of imagery data taken on two different dates can significantly improve the precision of crop area estimates for some crops.

Amis et al. (1981) investigated an alternative procedure to replace the clustering algorithm used in the original USDA EDITOR software. This alternative clustering procedure, called the CLASSY Clustering Algorithm, was originally proposed by Lennington and Malek (1978), and Lennington and Rassbach (1978, 1979a, 1979b). The algorithm is fundamentally a density estimation algorithm which approximates the overall data distribution as a mixture of multivariate normal distributions. The USDA EDITOR software is developed to train the pixel classifier described in Sigman et al. (1978). In Amis et al. (1981), the sample was treated as a statistically independent unlabeled sample $\{\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n\}$. In this case, the likelihood function becomes

$$L(\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) = \prod_{j=1}^{n} \left[\sum_{i=1}^{m} q_i p_i(\underset{\sim}{x}_j | \underset{\sim}{\mu}_i, \underset{\sim}{\Sigma}_i) \right] ,$$

where

q_i = a priori probability of occurrence of class i ,

$$p_i(\underset{\sim}{x}|\mu_i, \underset{\sim}{\Sigma}_i) = (2\pi)^{-p/2}|\underset{\sim}{\Sigma}_i|^{-1/2}\exp(-1/2\,(\underset{\sim}{x} - \mu_i)'\underset{\sim}{\Sigma}_i^{-1}(\underset{\sim}{x} - \mu_i)) ,$$

m = total number of classes .

In CLASSY, an a priori probability distribution α is imposed on the parameters $m, q_1,\ldots, q_m, \mu_1,\ldots, \mu_m, \underset{\sim}{\Sigma}_1,\ldots, \underset{\sim}{\Sigma}_n$. The discrete parameter m and the continuous parameters $q_1,\ldots, q_m, \mu_1,\ldots, \mu_m, \underset{\sim}{\Sigma}_1,\ldots, \underset{\sim}{\Sigma}_m$ are determined to maximize

$$R(\underset{\sim}{x}_1,\ldots, \underset{\sim}{x}_n) = \alpha\, L(\underset{\sim}{x}_1,\ldots, \underset{\sim}{x}_n) .$$

One advantage of this procedure is that CLASSY requires no decisions from the analyst concerning the number of clusters. In their study of northwest Missouri data, the CLASSY clustering algorithm was demonstrated to produce significantly better results in the sense of a higher R^2 value than the USDA EDITOR procedure when testing and training were done on the whole sample.

III. PROBLEM AND ANALYSES

A. Definition of the Problem

We consider a finite population of N clusters. Each cluster is a sampling unit consisting of a group of smaller units that we call elements or subunits. Let cluster i contain M_i subunits for $i = 1,2,\ldots, N$. Each element is placed in exactly one of c categories, where c is a known constant. A sample of n clusters is drawn from this population by the method of simple random sampling. Assume that the value of a vector of p auxiliary variables is available for each element of the population of N clusters. Let the variable θ assume the values $1,2,\ldots, c$; let θ_{ij} be observed for each element where $\theta_{ij} = k$ if the element belongs to category k, $k \in \{1,2,\ldots, c\}$. For each k, let

$$Y_{kij} = \begin{cases} 1 & \text{if element } j \text{ in cluster } i \text{ is placed in category } k \text{ ; i.e., } \theta_{ij} = k \\ 0 & \text{otherwise .} \end{cases} \qquad (3.A.1)$$

Let

$\underset{\sim}{X}_{ij}$ = a column vector of p auxiliary variables for element j in cluster i, for $i = 1,2,\ldots, N$, $j = 1,2,\ldots, M_i$. (3.A.2)

Then,

$$Y_{ki.} = \sum_{j=1}^{M_i} Y_{kij} = \text{number of elements belonging to category } k \text{ in cluster } i \text{ ,} \tag{3.A.3}$$

$$Y_{k..} = \sum_{i=1}^{N} Y_{ki.} = \text{the population Y-total for category } k$$

$$= \text{number of elements belonging to category } k \text{ in the population .}$$

In a sample cluster, the category to which an element belongs is observed. For convenience, let the first n population clusters be sampled. Hence, in sample cluster i $(i = 1,2,\ldots, n)$, $Y_{ki.}$ is known for $k = 1,2,\ldots, c$.

The problem of interest is to construct an estimator of the population Y-total for each category by using the auxiliary information contained in $\underset{\sim}{X}$.

B. Regression Estimation

In this section, we shall construct the regression estimator of the population Y-total for each category by employing the posterior probability as an auxiliary variable. As previously noted, for each chosen element, the variable Y_{kij} assumes values 0 and 1 . If we

call $Y = 1$ a 'success' and $Y = 0$ a 'failure', then the expected value of Y is the probability of success $p(Y = 1)$. In sampling theory, the linear regression estimator is designed to increase the precision of estimation by taking advantage of the linear correlation between the primary variable and the auxiliary variables. Therefore, it is necessary to assess the relation between the probability of success and the auxiliary vector $\underset{\sim}{X}$. Quite often a regression-like model will be constructed for the conditional expectation of Y given the auxiliary vector $\underset{\sim}{X}$, denoted by $p(Y = 1 \mid \underset{\sim}{X})$. The true conditional probability should be the "best" possible auxiliary variable.

We begin by assuming that for each secondary unit, the vectors $(\theta_{ij}, \underset{\sim}{X}'_{ij})'$, $j = 1,2,\ldots, M_i$, $i = 1,2,\ldots, N$, are generated from some multivariate distribution. Assume that the vector $\underset{\sim}{X}$, conditional on $\theta = k$, $(k = 1,2,\ldots, c)$ is distributed as a multivariate normal with mean $\underset{\sim}{\mu}_k$ and variance-covariance matrix $\underset{\sim}{\Sigma}_k$. That is, for each k, the probability density of $\underset{\sim}{X}$ is

$$p(\underset{\sim}{X} \mid \theta=k) = (2\pi)^{-p/2} \left|\underset{\sim}{\Sigma}_k\right|^{-1/2} \exp[-1/2\,(\underset{\sim}{X} - \underset{\sim}{\mu}_k)'\underset{\sim}{\Sigma}_k^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}_k)] .$$

Let f_k denote the probability that $\theta = k$ for category k $(k = 1,2,\ldots, c)$; i.e., $f_k = p(\theta=k)$. Assume that f_k is positive for each k. Then, the posterior probability that a secondary element with a value $\underset{\sim}{X}$ is from category k is

$$p(\theta=k\,|\,\underset{\sim}{X}) = \frac{p(\underset{\sim}{X}\,|\,\theta=k)f_k}{\sum_{i=1}^{c} p(\underset{\sim}{X}\,|\,\theta=i)f_i}\ , \quad k = 1,2,\ldots, c\ . \qquad (3.B.1)$$

We now develop estimators for the Y-total based upon the conditional probability $p(\theta=k\,|\,\underset{\sim}{X})$.

Case 1: All Parameters Known

If $\underset{\sim}{\mu}_i$, $\underset{\sim}{\Sigma}_i$, and f_i are all known, then the posterior probabilities can be used in the construction of a regression estimator of the population Y-total for each category. To achieve this construction, the sum of the conditional probabilities is created for each cluster. A regression estimator of $Y_{k..}$ is

$$\hat{Y}_{k..(\ell r)} = N\,\hat{\bar{Y}}_{k..(\ell r)}\ , \quad k = 1,2,\ldots, c\ ,$$

where

$$\hat{\bar{Y}}_{k..(\ell r)} = \bar{Y}^{(n)}_{k..} + \hat{\beta}_k(\bar{Z}^{(N)}_{k..} - \bar{Z}^{(n)}_{k..})\ , \qquad (3.B.2)$$

$$Z_{ki.} = \sum_{j=1}^{M_i} p(\theta=k\,|\,\underset{\sim}{X}_{ij})\ , \qquad (3.B.2.1)$$

$$\bar{Y}^{(n)}_{k..} = \frac{1}{n}\sum_{j=1}^{n} Y_{kj.}\ , \qquad (3.B.2.2)$$

$$\bar{Z}^{(N)}_{k..} = \frac{1}{N}\sum_{j=1}^{N} Z_{kj.}\ , \qquad (3.B.2.3)$$

$$\bar{Z}_{k..}^{(n)} = \frac{1}{n} \sum_{j=1}^{n} Z_{kj.} \,, \qquad (3.B.2.4)$$

$$\hat{\beta}_k = \frac{\sum_{j=1}^{n} (Y_{kj.} - \bar{Y}_{k..}^{(n)})(Z_{kj.} - \bar{Z}_{k..}^{(n)})}{\sum_{j=1}^{n} (Z_{kj.} - \bar{Z}_{k..}^{(n)})^2} \,. \qquad (3.B.2.5)$$

Note that $\hat{\beta}_k$ is the sample regression coefficient obtained in a regression of $Y_{ki.}$ on $Z_{ki.}$ with intercept.

Case 2: All Parameters Unknown

In the case where $\underset{\sim}{\mu}_i$, $\underset{\sim}{\Sigma}_i$, and f_i are all unknown, the normal conditional distribution of $\underset{\sim}{X}$ conditional on θ can be estimated by estimating the mean and covariance matrix for each category. Assume that for each k, the estimator $(\hat{\underset{\sim}{\mu}}_k, \hat{\underset{\sim}{\Sigma}}_k, \hat{f}_k)$ is available for $(\underset{\sim}{\mu}_k, \underset{\sim}{\Sigma}_k, f_k)$. Then, for each k, $p(\underset{\sim}{X}|\theta=k)$ can be estimated by

$$\hat{p}(\underset{\sim}{X}|\theta=k) = (2\pi)^{-p/2} |\hat{\underset{\sim}{\Sigma}}_k|^{-1/2} \exp[-1/2 (\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_k)' \hat{\underset{\sim}{\Sigma}}_k^{-1} (\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_k)] \,.$$

And for each k, the posterior probability that an element with a value $\underset{\sim}{X}$ is from category k is estimated by

$$\hat{p}(\theta=k|\underset{\sim}{X}) = \frac{\hat{p}(\underset{\sim}{X}|\theta=k)\hat{f}_k}{\sum_{i=1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i} \,, \quad k = 1,2,\ldots, c \,. \qquad (3.B.3)$$

The sum of the estimated conditional probabilities is created for each cluster to construct a regression estimator of the population Y-total for each category. The regression estimator of $Y_{k..}$ is

$$\tilde{Y}_{k..(\ell r)} = N \tilde{\bar{Y}}_{k..(\ell r)} , \quad k = 1,2,\ldots, c ,$$

where

$$\tilde{\bar{Y}}_{k..(\ell r)} = \bar{Y}^{(n)}_{k..} + \tilde{\beta}_k(\bar{W}^{(N)}_{k..} - \bar{W}^{(n)}_{k..}) , \tag{3.B.4}$$

$$W_{ki.} = \sum_{j=1}^{M_i} \hat{p}(\theta=k|\underset{\sim}{X}_{ij}) ,$$

$$\bar{Y}^{(n)}_{k..} = \frac{1}{n} \sum_{j=1}^{n} Y_{kj.} ,$$

$$\bar{W}^{(N)}_{k..} = \frac{1}{N} \sum_{j=1}^{N} W_{kj.} ,$$

$$W^{(n)}_{k..} = \frac{1}{n} \sum_{j=1}^{n} W_{kj.} ,$$

$$\tilde{\beta}_k = \frac{\sum_{j=1}^{n} (Y_{kj.} - \bar{Y}^{(n)}_{k..})(W_{kj.} - \bar{W}^{(n)}_{k..})}{\sum_{j=1}^{n} (W_{kj.} - \bar{W}^{(n)}_{k..})^2} .$$

One can see that $\tilde{\beta}_k$ is the sample regression coefficient obtained in a regression of $Y_{ki.}$ on $W_{ki.}$ with intercept.

Note that this regression estimator should have larger variance than the regression estimator of Case 1 because the explanatory variable is estimated. The asymptotic properties of the estimated posterior probability will be investigated in Section D.

C. Basic Definitions and Results

In this section, we summarize some well-known definitions and results of matrix algebra. These definitions and results are useful for cases where it is convenient to arrange the elements of the matrix as a vector. These definitions and results are presented in Fuller (1981) and in Henderson and Searle (1979).

Definition 3.C.1. Let $\underset{\sim}{A} = (a_{ij})$ be a $r \times s$ matrix. Then

$$\text{vec } \underset{\sim}{A} = (a_{11}, a_{21}, \ldots, a_{r1}, a_{12}, a_{22}, \ldots, a_{r2}, \ldots, a_{1s}, a_{2s}, \ldots, a_{rs})' .$$

Definition 3.C.2. Let $\underset{\sim}{A} = (a_{ij})$ be a $r \times r$ matrix. Then

$$\text{vech } \underset{\sim}{A} = (a_{11}, a_{21}, \ldots, a_{r1}, a_{22}, \ldots, a_{r2}, a_{33}, a_{43}, \ldots, a_{r3}, \ldots, a_{rr})' .$$

For any symmetric matrix, $\underset{\sim}{A} = (a_{ij})$ of order r, vec $\underset{\sim}{A}$ and vech $\underset{\sim}{A}$ are linear transformations of one another. We represent these transformations by the matrices $\underset{\sim}{\Phi}$ and $\underset{\sim}{\Psi}$. Let

$$\text{vec } \underset{\sim}{A} = \underset{\sim}{\Phi} \text{ vech } \underset{\sim}{A} ,$$

and

$$\text{vech } \underset{\sim}{A} = \underset{\sim}{\Psi} \text{ vec } \underset{\sim}{A} .$$

Note that $\underset{\sim}{\Phi}$ is a unique $p^2 \times \frac{1}{2} p(p+1)$ matrix and of full column rank. However, there are many transformations of vec $\underset{\sim}{A}$ into vech $\underset{\sim}{A}$. Among them the Moore-Penrose generalized inverse $(\underset{\sim}{\Phi}'\underset{\sim}{\Phi})^{-1}\underset{\sim}{\Phi}'$ of $\underset{\sim}{\Phi}$ is particularly useful. The product $\underset{\sim}{\Psi}\underset{\sim}{\Phi}$ is an identity matrix.

Definition 3.C.3. Let $\underset{\sim}{A} = (a_{ij})$ be a $p \times q$ matrix and $\underset{\sim}{B} = (b_{ij})$ be a $r \times s$ matrix. The Kronecker product of $\underset{\sim}{A}$ and $\underset{\sim}{B}$ is the $pr \times qs$ matrix given by

$$\underset{\sim}{A} \otimes \underset{\sim}{B} = \begin{pmatrix} a_{11}\underset{\sim}{B} & a_{12}\underset{\sim}{B} & \cdots & a_{1q}\underset{\sim}{B} \\ a_{21}\underset{\sim}{B} & a_{22}\underset{\sim}{B} & \cdots & a_{2q}\underset{\sim}{B} \\ & & & \\ a_{p1}\underset{\sim}{B} & a_{p2}\underset{\sim}{B} & \cdots & a_{pq}\underset{\sim}{B} \end{pmatrix} .$$

Definition 3.C.4. For any matrix $\underset{\sim}{A} = (a_{ij})$, the total differential of $\underset{\sim}{A}$, denoted by $d\underset{\sim}{A}$, is given by

$$d\underset{\sim}{A} = (d\, a_{ij}) .$$

Definition 3.C.5. Let $\underset{\sim}{a} = \underset{\sim}{a}(\underset{\sim}{\theta})$ be a p-dimensional column vector whose typical element $a_j = a_j(\underset{\sim}{\theta})$ is a function of the r-dimensional column vector $\underset{\sim}{\theta}$. Then,

$$\frac{\partial \underset{\sim}{a}}{\partial \underset{\sim}{\theta}} = \begin{pmatrix} \frac{\partial a_1}{\partial \theta_1} & \frac{\partial a_1}{\partial \theta_2} & \cdots & \frac{\partial a_1}{\partial \theta_q} \\ \frac{\partial a_2}{\partial \theta_1} & \frac{\partial a_2}{\partial \theta_2} & \cdots & \frac{\partial a_2}{\partial \theta_q} \\ \vdots & \vdots & & \\ \frac{\partial a_p}{\partial \theta_1} & \frac{\partial a_p}{\partial \theta_2} & \cdots & \frac{\partial a_p}{\partial \theta_q} \end{pmatrix} .$$

Definition 3.C.6. Let $g(\underset{\sim}{A})$ be a scalar function of the $p \times q$ matrix $\underset{\sim}{A} = (a_{ij})$. Then, $\frac{\partial g(\underset{\sim}{A})}{\partial \underset{\sim}{A}}$ is the $p \times q$ matrix with the ij-th element being $\frac{\partial g(\underset{\sim}{A})}{\partial a_{ij}}$.

Result 3.C.1. Let the matrices $\underset{\sim}{A}$, $\underset{\sim}{B}$, $\underset{\sim}{C}$, and $\underset{\sim}{D}$ be suitably comformable matrices. Then,

(i) $(\underset{\sim}{A} \boxtimes \underset{\sim}{B})(\underset{\sim}{C} \boxtimes \underset{\sim}{D}) = (\underset{\sim}{AC}) \boxtimes (\underset{\sim}{BD})$,

(ii) $(\underset{\sim}{A} \boxtimes \underset{\sim}{B})' = \underset{\sim}{A}' \boxtimes \underset{\sim}{B}'$,

(iii) $(\underset{\sim}{A} \boxtimes \underset{\sim}{B})^{-1} = \underset{\sim}{A}^{-1} \boxtimes \underset{\sim}{B}^{-1}$,

(iv) $(\underset{\sim}{A} + \underset{\sim}{B}) \boxtimes (\underset{\sim}{C} + \underset{\sim}{D}) = (\underset{\sim}{A} \boxtimes \underset{\sim}{C}) + (\underset{\sim}{A} \boxtimes \underset{\sim}{D}) + (\underset{\sim}{B} \boxtimes \underset{\sim}{C}) + (\underset{\sim}{B} \boxtimes \underset{\sim}{D})$.

Result 3.C.2. Let $\underset{\sim}{A}$, $\underset{\sim}{B}$, and $\underset{\sim}{C}$ be $p \times q$, $q \times m$, and $m \times n$ matrices, respectively. Then,

$$\text{vec}(\underset{\sim}{ABC}) = (\underset{\sim}{C}' \otimes \underset{\sim}{A})\text{vec } \underset{\sim}{B} .$$

Proof. [See Fuller (1981).]

Result 3.C.3. Let $\underset{\sim}{A}$ and $\underset{\sim}{B}$ be $p \times q$ and $q \times p$ matrices, respectively, and let the trace of $\underset{\sim}{C}$, denoted by $\text{tr } \underset{\sim}{C}$, be the sum of the diagonal elements of a square matrix $\underset{\sim}{C}$. Then,

$$\text{tr}(\underset{\sim}{AB}) = (\text{vec } \underset{\sim}{A}')'\text{vec } \underset{\sim}{B} = (\text{vec } \underset{\sim}{B}')'\text{vec } \underset{\sim}{A} .$$

Proof. [See Fuller (1981).]

Result 3.C.4. [Neudecker (1969)] (i) Let $\underset{\sim}{A}$ and $\underset{\sim}{B}$ be conformable matrices. Then, $d(\underset{\sim}{A}\ \underset{\sim}{B}) = (d\ \underset{\sim}{A})\underset{\sim}{B} + \underset{\sim}{A}(d\ \underset{\sim}{B})$.

(ii) If ℓ is a linear scalar function of the matrix $\underset{\sim}{A}$, then

$$d(\ell\underset{\sim}{A}) = \ell(d\ \underset{\sim}{A}) .$$

(iii) $d(\text{vec } \underset{\sim}{A}) = \text{vec } (d\ \underset{\sim}{A})$.

Result 3.C.5. [Neudecker (1969)] Let $\underset{\sim}{y}$ be a column vector of which every element is a function of the column vector $\underset{\sim}{x}$ such that $d\ \underset{\sim}{y} = \underset{\sim}{M}(d\ \underset{\sim}{x})$ for some matrix $\underset{\sim}{M}$. Then,

$$\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}} = \underset{\sim}{M}' \ .$$

Result 3.C.6. Let $\underset{\sim}{A}$, $\underset{\sim}{B}$, and $\underset{\sim}{H}$ be suitably comformable matrices such that $d\,\underset{\sim}{G} = \underset{\sim}{A}(d\,\underset{\sim}{H})\underset{\sim}{B}$. Then,

$$\frac{\partial \text{ vec } \underset{\sim}{G}}{\partial \text{ vec } \underset{\sim}{H}} = \underset{\sim}{B} \boxtimes \underset{\sim}{A}' \ ,$$

and

$$\frac{\partial \text{ vech } \underset{\sim}{G}}{\partial \text{ vech } \underset{\sim}{H}} = \underset{\sim}{\Phi}'(\underset{\sim}{B} \boxtimes \underset{\sim}{A}')\underset{\sim}{\Psi}' \ .$$

Proof. $d(\text{vec } \underset{\sim}{G}) = \text{vec}(d\,\underset{\sim}{G}) = (\underset{\sim}{B}' \boxtimes \underset{\sim}{A})\text{vec}(d\,\underset{\sim}{H}) = (\underset{\sim}{B}' \boxtimes \underset{\sim}{A})d(\text{vec } \underset{\sim}{H})$. Also,

$$d(\text{vech } \underset{\sim}{G}) = \underset{\sim}{\Psi}[d(\text{vec } \underset{\sim}{G})] = \underset{\sim}{\Psi}(\underset{\sim}{B}' \boxtimes \underset{\sim}{A})\text{vec}(d\,\underset{\sim}{H})$$

$$= \underset{\sim}{\Psi}(\underset{\sim}{B}' \boxtimes \underset{\sim}{A})\underset{\sim}{\Phi}[d(\text{vech } \underset{\sim}{H})] \ . \qquad \square$$

Result 3.C.7. Let $\underset{\sim}{A}$ be a symmetric nonsingular matrix. Then,

$$\frac{\partial \text{ vec } \underset{\sim}{A}^{-1}}{\partial \text{ vec } \underset{\sim}{A}} = -\,(\underset{\sim}{A} \boxtimes \underset{\sim}{A})^{-1}$$

and

$$\frac{\partial \text{ vech } \underset{\sim}{A}^{-1}}{\partial \text{ vech } \underset{\sim}{A}} = - \underset{\sim}{\Psi}'(\underset{\sim}{A} \boxtimes \underset{\sim}{A})^{-1} \underset{\sim}{\Psi}' \ .$$

Proof. Since $\underset{\sim}{A}\,\underset{\sim}{A}^{-1} = \underset{\sim}{I}$,

$$\underset{\sim}{0} = d\,\underset{\sim}{I} = d(\underset{\sim}{A}\underset{\sim}{A}^{-1}) = \underset{\sim}{A}(d\,\underset{\sim}{A}^{-1}) + (d\,\underset{\sim}{A})\underset{\sim}{A}^{-1} \ .$$

Thus,

$$d\,\underset{\sim}{A}^{-1} = - \underset{\sim}{A}^{-1}(d\,\underset{\sim}{A})\underset{\sim}{A}^{-1} \ .$$

Using Result 3.C.6, we have

$$\frac{\partial \text{ vec } \underset{\sim}{A}^{-1}}{\partial \text{ vec } \underset{\sim}{A}} = - (\underset{\sim}{A} \boxtimes \underset{\sim}{A})^{-1} \ ,$$

and

$$\frac{\partial \text{ vech } \underset{\sim}{A}^{-1}}{\partial \text{ vech } \underset{\sim}{A}} = - \underset{\sim}{\Psi}'(\underset{\sim}{A} \boxtimes \underset{\sim}{A})^{-1} \underset{\sim}{\Psi}' \ . \qquad \square$$

Result 3.C.8. Let $\underset{\sim}{A}$ be a symmetric nonsingular matrix. Then,

$$\frac{\partial |\underset{\sim}{A}|}{\partial \text{ vech } \underset{\sim}{A}} = |\underset{\sim}{A}| \text{vech}[2\underset{\sim}{A}^{-1} - \text{diag}(\underset{\sim}{A}^{-1})] \ ,$$

where $\text{diag}(\underset{\sim}{A}^{-1})$ is the diagonal matrix of $\underset{\sim}{A}^{-1}$.

Proof. [See Fuller (1981).]

Result 3.C.9. Let $\underset{\sim}{Z}_t$ be normally distributed with mean $\underset{\sim}{\mu}_Z$ and covariance matrix $\underset{\sim}{\Sigma}_{ZZ}$. Let

$$\underset{\sim}{m}_{ZZ} = (n-1)^{-1} \sum_{t=1}^{n} (\underset{\sim}{Z}_t - \bar{\underset{\sim}{Z}})'(\underset{\sim}{Z}_t - \bar{\underset{\sim}{Z}}) .$$

Then the covariance matrix of vech $\underset{\sim}{m}_{ZZ}$ is

$$V(\text{vech } \underset{\sim}{m}_{ZZ}) = 2(n-1)^{-1} \underset{\sim}{\psi}(\underset{\sim}{\Sigma}_{ZZ} \otimes \underset{\sim}{\Sigma}_{ZZ})\underset{\sim}{\psi}' .$$

Proof. [See Fuller (1981).]

D. Asymptotic Properties of the Estimated Posterior Probability

In this section we suppose that the finite population is a sample from a multivariate infinite population. This superpopulation assumption is imposed so that we may investigate the limiting properties of the estimators. We first establish some properties of the estimated conditional probabilities.

Without loss of generality, let us consider $\hat{p}(\theta=1|\underset{\sim}{X})$, where

$$\hat{p}(\theta=1|\underset{\sim}{X}) = \frac{\hat{p}(\underset{\sim}{X}|\theta=1)\hat{f}_1}{\sum_{i=1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i} ,$$

$$\hat{p}(\underset{\sim}{X}|\theta=i) = (2\pi)^{-p/2}|\hat{\underset{\sim}{\Sigma}}_i|^{-1/2} \exp\{-1/2 \operatorname{tr}[\hat{\underset{\sim}{\Sigma}}_i^{-1}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_i)(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_i)']\} .$$

Define, for $j = 1,2,\ldots, c$,

$$\hat{\underset{\sim}{A}}_j = (\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_j)(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_j)' ,$$

$$\underset{\sim}{A}_j = (\underset{\sim}{X} - \underset{\sim}{\mu}_j)(\underset{\sim}{X} - \underset{\sim}{\mu}_j)' ,$$

$$\underset{\sim}{b} = (\hat{\underset{\sim}{\mu}}_1', \ldots, \hat{\underset{\sim}{\mu}}_c', (\text{vech } \hat{\underset{\sim}{\Sigma}}_1)', \ldots, (\text{vech } \hat{\underset{\sim}{\Sigma}}_c)', \hat{f}_1, \ldots, \hat{f}_c)' ,$$

$$\underset{\sim}{b}_0 = (\underset{\sim}{\mu}_1', \ldots, \underset{\sim}{\mu}_c', (\text{vech } \underset{\sim}{\Sigma}_1)', \ldots, (\text{vech } \underset{\sim}{\Sigma}_c)', f_1, \ldots, f_c)' .$$

Also, for any function $h(\gamma)$, denote by $\frac{\partial\, h(\gamma_0)}{\partial\, \gamma}$ the derivative $\frac{\partial\, h(\gamma)}{\partial\, \gamma}$ evaluated at $\gamma = \gamma_0$.

Now, for a fixed $\underset{\sim}{X}$, $\hat{p}(\theta=1|\underset{\sim}{X})$ is a continuous function of $\underset{\sim}{b}$ and has continuous partial derivatives of all orders at $\underset{\sim}{b}_0$. Differentiating $\hat{p}(\underset{\sim}{X}|\theta=1)$ with respect to $\underset{\sim}{b}$, we have

$$\frac{\partial\, \hat{p}(\underset{\sim}{X}|\theta=1)}{\partial\, \hat{\underset{\sim}{\mu}}_1} = (2\pi)^{-p/2} |\hat{\underset{\sim}{\Sigma}}_1|^{-1/2} \exp[-1/2\,(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)' \hat{\underset{\sim}{\Sigma}}_1^{-1} (\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)] \hat{\underset{\sim}{\Sigma}}_1^{-1} (\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)$$

$$= \eta(\hat{\underset{\sim}{\mu}}_1, \hat{\underset{\sim}{\Sigma}}_1|\underset{\sim}{X}) [\hat{\underset{\sim}{\Sigma}}_1^{-1} (\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)] ,$$

$$\frac{\partial\, \hat{p}(\underset{\sim}{X}|\theta=1)}{\partial\, \text{vech}\, \hat{\underset{\sim}{\Sigma}}_1} = (2\pi)^{-p/2}(-\tfrac{1}{2})\left|\hat{\underset{\sim}{\Sigma}}_1\right|^{-\frac{1}{2}} [2\text{vech}\, \hat{\underset{\sim}{\Sigma}}_1^{-1}$$

$$- \text{vech diag}(\hat{\underset{\sim}{\Sigma}}_1^{-1})]\exp[-\tfrac{1}{2}(\underset{\sim}{X}$$

$$- \hat{\underset{\sim}{\mu}}_1)'\hat{\underset{\sim}{\Sigma}}_1^{-1}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)]$$

$$+ (2\pi)^{-p/2}\left|\hat{\underset{\sim}{\Sigma}}_1\right|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)'\hat{\underset{\sim}{\Sigma}}_1^{-1}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)]$$

$$\times \frac{1}{2}\underset{\sim}{\Phi}'(\hat{\underset{\sim}{\Sigma}}_1^{-1} \boxtimes \hat{\underset{\sim}{\Sigma}}_1^{-1})\underset{\sim}{\Psi}'\underset{\sim}{\Phi}'\underset{\sim}{\Phi}\, \text{vech}\, \hat{\underset{\sim}{A}}_1$$

$$= \frac{1}{2}\eta(\hat{\underset{\sim}{\mu}}_1, \hat{\underset{\sim}{\Sigma}}_1|\underset{\sim}{X})[\underset{\sim}{\Phi}'(\hat{\underset{\sim}{\Sigma}}_1^{-1} \boxtimes \hat{\underset{\sim}{\Sigma}}_1^{-1})\underset{\sim}{\Phi}\, \text{vech}\, \hat{\underset{\sim}{A}}_1$$

$$- \text{vech}[2\hat{\underset{\sim}{\Sigma}}_1^{-1} - \text{diag}(\hat{\underset{\sim}{\Sigma}}_1^{-1})]\} ,$$

where

$$\eta(\hat{\underset{\sim}{\mu}}_1, \hat{\underset{\sim}{\Sigma}}_1|\underset{\sim}{X}) = (2\pi)^{-p/2}\left|\hat{\underset{\sim}{\Sigma}}_1\right|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)'\hat{\underset{\sim}{\Sigma}}_1^{-1}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)] .$$

For $j \neq 1$,

$$\frac{\partial\, \hat{p}(\underset{\sim}{X}|\theta=1)}{\partial\, \hat{\underset{\sim}{\mu}}_j} = 0 ,$$

and

$$\frac{\partial\, \hat{p}(\underset{\sim}{X}|\theta=1)}{\partial\, \text{vech}\, \hat{\underset{\sim}{\Sigma}}_j} = 0 .$$

Thus, we have

$$\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \hat{\underset{\sim}{\mu}}_1} = \hat{f}_1\Big(\sum_{i=1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big)^{-2}\Big(\sum_{i\neq 1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big) \frac{\partial \hat{p}(\underset{\sim}{X}|\theta=1)}{\partial \hat{\underset{\sim}{\mu}}_1}$$

$$= \hat{p}(\theta=1|\underset{\sim}{X})[1 - \hat{p}(\theta=1|\underset{\sim}{X})][\hat{\underset{\sim}{\Sigma}}_1^{-1}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_1)] , \qquad (3.D.1)$$

$$\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \text{ vech } \hat{\underset{\sim}{\Sigma}}_1} = \hat{f}_1\Big(\sum_{i=1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big)^{-2}\Big(\sum_{i\neq 1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big) \frac{\partial \hat{p}(\underset{\sim}{X}|\theta=1)}{\partial \text{ vech } \hat{\underset{\sim}{\Sigma}}_1} ,$$

$$= \tfrac{1}{2}\hat{p}(\theta=1|\underset{\sim}{X})[1 - \hat{p}(\theta=1|\underset{\sim}{X})]\{\underset{\sim}{\Phi}'(\hat{\underset{\sim}{\Sigma}}_1^{-1} \otimes \hat{\underset{\sim}{\Sigma}}_1^{-1})\underset{\sim}{\Psi} \text{ vech } \hat{\underset{\sim}{A}}_1$$

$$- 2\text{vech } \hat{\underset{\sim}{\Sigma}}_1^{-1} + \text{vech}[\text{diag}(\hat{\underset{\sim}{\Sigma}}_1^{-1})]\} , \qquad (3.D.2)$$

$$\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \hat{f}_1} = \hat{p}(\underset{\sim}{X}|\theta=1)\Big(\sum_{i=1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big)^{-2}\Big(\sum_{i\neq 1}^{c} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big)$$

$$= \hat{f}_1^{-1}\hat{p}(\theta=1|\underset{\sim}{X})[1 - \hat{p}(\theta=1|\underset{\sim}{X})] . \qquad (3.D.3)$$

For $j \neq 1$,

$$\frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, \hat{\underset{\sim}{\mu}}_j} = -\hat{p}(\underset{\sim}{X}|\theta=1)\hat{f}_1\hat{f}_j\Big(\sum_{i=1}^{c}\hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big)^{-2}\frac{\partial\, \hat{p}(\underset{\sim}{X}|\theta=j)}{\partial\, \hat{\underset{\sim}{\mu}}_j}$$

$$= -\hat{p}(\theta=1|\underset{\sim}{X})\hat{p}(\theta=j|\underset{\sim}{X})[\hat{\underset{\sim}{\Sigma}}_j^{-1}(\underset{\sim}{X} - \hat{\underset{\sim}{\mu}}_j)] \ , \qquad (3.D.4)$$

$$\frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, \text{vech}\, \hat{\underset{\sim}{\Sigma}}_j} = -\hat{p}(\underset{\sim}{X}|\theta=1)\hat{f}_1\hat{f}_j\Big(\sum_{i=1}^{c}\hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big)^{-2}\frac{\partial\, \hat{p}(\underset{\sim}{X}|\theta=j)}{\partial\, \text{vech}\, \hat{\underset{\sim}{\Sigma}}_j}$$

$$= -1/2\, \hat{p}(\theta=1|\underset{\sim}{X})\hat{p}(\theta=j|\underset{\sim}{X})\{\underset{\sim}{D}'(\hat{\underset{\sim}{\Sigma}}_j^{-1} \otimes \hat{\underset{\sim}{\Sigma}}_j^{-1})\underset{\sim}{D}\, \text{vech}\, \hat{\underset{\sim}{A}}_j$$

$$- 2\text{vech}\, \hat{\underset{\sim}{\Sigma}}_j^{-1} + \text{vech}[\text{diag}(\hat{\underset{\sim}{\Sigma}}_j^{-1})]\} \ , \qquad (3.D.5)$$

and

$$\frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, \hat{f}_j} = -\hat{p}(\underset{\sim}{X}|\theta=1)\hat{f}_1\Big(\sum_{i=1}^{c}\hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i\Big)^{-2}\hat{p}(\underset{\sim}{X}|\theta=j)$$

$$= -\hat{f}_j^{-1}\hat{p}(\theta=1|\underset{\sim}{X})\hat{p}(\theta=j|\underset{\sim}{X}) \ . \qquad (3.D.6)$$

In the following theorem, we demonstrate the order of the error of $\hat{p}(\theta=1|\underset{\sim}{X})$ for a fixed $\underset{\sim}{X}$ given estimators of the parameters that have error whose order is $n^{-1/2}$.

Theorem 3.D.1. Assume that for $i = 1,2,\ldots, c$,

$$\hat{\mu}_i = \mu_i + O_p(n^{-1/2}) ,$$

$$\hat{\Sigma}_i = \Sigma_i + O_p(n^{-1/2}) ,$$

$$\hat{f}_i = f_i + O_p(n^{-1/2}) ,$$

where n is the number of sampled clusters. Then, for a fixed $\underset{\sim}{X}$,

$$\hat{p}(\theta=1|\underset{\sim}{X}) = p(\theta=1|\underset{\sim}{X}) + O_p(n^{-1/2}) ,$$

where $p(\theta=1|\underset{\sim}{X})$ and $\hat{p}(\theta=1|\underset{\sim}{X})$ are defined in (3.B.1) and (3.B.3), respectively.

Proof. Given $\underset{\sim}{X}$, a Taylor's series expansion of $\hat{p}(\theta=1|\underset{\sim}{X})$ about $\underset{\sim}{b}_0$ gives

$$\hat{p}(\theta=1|\underset{\sim}{X}) = p(\theta=1|\underset{\sim}{X}) + \sum_{i=1}^{c} \left(\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \underset{\sim}{\mu}_i^*} \right)' (\hat{\mu}_i - \mu_i)$$

$$+ \sum_{j=1}^{c} \left(\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \operatorname{vech} \underset{\sim}{\Sigma}_j^*} \right)' (\operatorname{vech} \hat{\underset{\sim}{\Sigma}}_j - \operatorname{vech} \underset{\sim}{\Sigma}_j)$$

$$+ \sum_{i=1}^{c} \frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial f_i^*} (\hat{f}_i - f_i) ,$$

where $\underset{\sim}{b}_0^* = (\underset{\sim}{\mu}_1^*, \ldots, \underset{\sim}{\mu}_c^*, \text{vech } \Sigma_1^*, \ldots, \text{vech } \Sigma_c^*, f_1^*, \ldots, f_c^*)$ is on the line joining $\underset{\sim}{b}$ and $\underset{\sim}{b}_0$.

Since the derivatives of $\hat{p}(\theta=1|\underset{\sim}{X})$ are continuous at $\underset{\sim}{b}_0$, for a fixed $\underset{\sim}{X}$,

$$\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \underset{\sim}{\mu}_i^*} = O_p(1) ,$$

$$\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \text{ vech } \underset{\sim}{\Sigma}_i^*} = O_p(1) ,$$

$$\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial f_i^*} = O_p(1) , \quad i = 1,2,\ldots, c .$$

By Lemma 5.1.4 in Fuller (1976),

$$\hat{p}(\theta=1|\underset{\sim}{X}) - p(\theta=1|\underset{\sim}{X}) = O_p(n^{-1/2}) . \qquad \square$$

The asymptotic properties of $\hat{p}(\theta=1|\underset{\sim}{X})$ for a fixed $\underset{\sim}{X}$ are given by the following theorem.

Theorem 3.D.2. Assume that, as $n \longrightarrow \infty$,

$$n^{1/2}(\underset{\sim}{b} - \underset{\sim}{b}_0) \xrightarrow{L} N(\underset{\sim}{0}, \underset{\sim}{V}) ,$$

where

$$\underset{\sim}{b} - \underset{\sim}{b}_0 = [(\hat{\mu}_1 - \mu_1)', \ldots, (\hat{\mu}_c - \mu_c)', (\text{vech}\{\hat{\underset{\sim}{\Sigma}}_1 - \underset{\sim}{\Sigma}_1\})', \ldots,$$

$$(\text{vech}\{\hat{\underset{\sim}{\Sigma}}_c - \underset{\sim}{\Sigma}_c\})' , (\hat{f}_1 - f_1), \ldots, (\hat{f}_c - f_c)]'$$

n is the number of sampled clusters,

$$\underset{\sim}{V} = \begin{pmatrix} \underset{\sim}{V}_{11} & \underset{\sim}{V}_{12} & \underset{\sim}{V}_{13} \\ \underset{\sim}{V}_{12}' & \underset{\sim}{V}_{22} & \underset{\sim}{V}_{23} \\ \underset{\sim}{V}_{13}' & \underset{\sim}{V}_{23}' & \underset{\sim}{V}_{33} \end{pmatrix} ,$$

$$\underset{\sim}{E}_1 = n^{1/2}[(\hat{\mu}_1 - \mu_1), \ldots, (\hat{\mu}_c - \mu_c)]' ,$$

$$\underset{\sim}{E}_2 = n^{1/2}[\text{vech}(\hat{\underset{\sim}{\Sigma}}_1 - \underset{\sim}{\Sigma}_1), \ldots, \text{vech}(\hat{\underset{\sim}{\Sigma}}_c - \underset{\sim}{\Sigma}_c)]' ,$$

$$\underset{\sim}{E}_3 = n^{1/2}[(\hat{f}_1 - f_1), \ldots, (\hat{f}_c - f_c)] ,$$

$$\underset{\sim}{E} = (\underset{\sim}{E}_1', \underset{\sim}{E}_2', \underset{\sim}{E}_3')' ,$$

$$\underset{\sim}{V}_{ij} = \mathrm{Cov}(\underset{\sim}{E}_i, \underset{\sim}{E}_j) \ , \quad i,j = 1,2,3 \ .$$

Then, as $n \longrightarrow \infty$, for a fixed $\underset{\sim}{X}$,

$$n^{1/2} [\hat{p}(\theta=1|\underset{\sim}{X}) - p(\theta=1|\underset{\sim}{X})] \xrightarrow{L} N(0, \underset{\sim}{B}\ \underset{\sim}{V}\ \underset{\sim}{B}') \ ,$$

where

$$\underset{\sim}{B}_1 = \left[\left(\frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, \mu_1} \right)', \ldots, \left(\frac{\partial\, \hat{p}(\underset{\sim}{\theta}=1|X)}{\partial\, \mu_c} \right)' \right],$$

$$\underset{\sim}{B}_2 = \left[\left(\frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, \mathrm{vech}\ \Sigma_1} \right)', \ldots, \left(\frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, \mathrm{vech}\ \Sigma_c} \right)' \right],$$

$$\underset{\sim}{B}_3 = \left[\frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, f_1}, \ldots, \frac{\partial\, \hat{p}(\theta=1|\underset{\sim}{X})}{\partial\, f_c} \right],$$

$$\underset{\sim}{B} = (\underset{\sim}{B}_1\ \underset{\sim}{B}_2\ \underset{\sim}{B}_3) \ ,$$

and the derivatives in $\underset{\sim}{B}$ are given in (3.D.1) - (3.D.6).

Proof. Given $\underset{\sim}{X}$, $\hat{p}(\theta=1|\underset{\sim}{X})$ is continuous at $\underset{\sim}{b} = \underset{\sim}{b}_0$. Using a Taylor series expansion of $\hat{p}(\theta=1|\underset{\sim}{X})$ about $\underset{\sim}{b}_0$, we have, for $\underset{\sim}{X}$ given,

$$\hat{p}(\theta=1|\underset{\sim}{X}) = p(\theta=1|\underset{\sim}{X}) + \sum_{i=1}^{c} \left(\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial \mu_i} \right)' (\hat{\mu}_i - \mu_i)$$

$$+ \sum_{j=1}^{c} \left(\frac{\partial \hat{p}(\underset{\sim}{\theta}=1|X)}{\partial \text{ vech } \underset{\sim}{\Sigma}_j} \right)' (\text{vech } \hat{\underset{\sim}{\Sigma}}_j - \text{vech } \underset{\sim}{\Sigma}_j)$$

$$+ \sum_{i=1}^{c} \left(\frac{\partial \hat{p}(\theta=1|\underset{\sim}{X})}{\partial f_i} \right) (\hat{f}_i - f_i)$$

$$+ O_p(n^{-1}) .$$

[See Fuller (1976).]

It follows that as $n \longrightarrow \infty$, for each i,

$$\text{plim}[n^{1/2} \{\hat{p}(\theta=1|\underset{\sim}{X}) - p(\theta=1|\underset{\sim}{X})\} - \underset{\sim}{B}\,\underset{\sim}{E}] = 0 .$$

Therefore, the limiting distribution of $n^{1/2}[\hat{p}(\theta=1|\underset{\sim}{X}) - p(\theta=1|\underset{\sim}{X})]$ is the same as the limiting distribution of $\underset{\sim}{B}\,\underset{\sim}{E}$. And hence,

$n^{1/2}\,[\hat{p}(\theta=1|\underset{\sim}{X}) - p(\theta=1|\underset{\sim}{X})]$ is asymptotically distributed as a normal vector with mean zero and covariance matrix $\underset{\sim}{B}\,V\,\underset{\sim}{B}'$. □

For a self-weighting sample of n clusters, the mean, covariance matrix, and category probability can be estimated by the sample mean, sample covariance matrix, and sample fraction for category k , respectively. That is, for each k ,

$$\hat{\underset{\sim}{\mu}}_k = \bar{\underset{\sim}{X}}_k = \Big[\sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij}\Big]^{-1} \sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij}\underset{\sim}{X}_{ij} \,, \tag{3.D.7}$$

$$\hat{\underset{\sim}{\Sigma}}_k = \underset{\sim}{S}_k = \Big[\sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij} - 1\Big]^{-1} \sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij}(\underset{\sim}{X}_{ij} - \hat{\underset{\sim}{\mu}}_k)(\underset{\sim}{X}_{ij} - \hat{\underset{\sim}{\mu}}_k)' \tag{3.D.8}$$

and

$$\hat{f}_k = \Big[\sum_{i=1}^{n} M_i\Big]^{-1} \sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij} \,. \tag{3.D.9}$$

These estimators are easily modified for designs that are not self-weighting. In the following theorem, we demonstrate the order of the error of the estimators $\hat{\underset{\sim}{\mu}}_k$, $\hat{\underset{\sim}{\Sigma}}_k$, $\hat{f}_k$, $k = 1,2,\ldots, c$.

Theorem 3.D.3. Let $\{\xi_n: n = 1,2,\ldots,\}$ be a sequence of finite populations, where ξ_n contains N_n clusters, $N_n > N_{n-1}$. Let M_i denote the number of secondary elements in cluster i for each i. A sample of n clusters is selected from population ξ_n by simple random sampling without replacement. Let

$$\lim_{n\to\infty} n N_n^{-1} = h ,$$

where $0 < h \leq 1$. Let

$$\underset{\sim}{X}_{ij} = \underset{\sim}{u}_i + \underset{\sim}{\varepsilon}_{ij} ,$$

$$Y_{kij} = v_{ki} + \eta_{kij} , \quad k = 1,2,\ldots, c , \quad j = 1,2,\ldots, M_i ,$$

$$i = 1,2,\ldots, N_n ,$$

where the random vector $(\underset{\sim}{u}_i', v_{1i},\ldots, v_{ci})$ is the 'primary component' and the random vector $(\underset{\sim}{\varepsilon}_{ij}', \eta_{1ij},\ldots, \eta_{cij})$ is the 'secondary component'. Let $\underset{\sim}{v}_i = (v_{1i},\ldots, v_{ci})'$ and $\underset{\sim}{\eta}_{ij} = (\eta_{1ij},\ldots, \eta_{cij})'$ for every i,j. We assume:

(i) For cluster i, the vectors $(\underset{\sim}{\varepsilon}_{ij}', \underset{\sim}{\eta}_{ij}')'$, $j = 1,2,\ldots, M_i$, are a random sample from an infinite population with zero mean vector, uniformly bounded fourth moments, and positive definite covariance matrix, $\underset{\sim}{\Sigma}_i$, where

$$\underset{\sim}{\Sigma}_i = \begin{pmatrix} \underset{\sim}{\Sigma}_{\varepsilon\varepsilon i} & \underset{\sim}{\Sigma}_{\varepsilon\eta i} \\ \underset{\sim}{\Sigma}'_{\varepsilon\eta i} & \underset{\sim}{\Sigma}_{\eta\eta i} \end{pmatrix} .$$

(ii) The vectors $(\underset{\sim}{u}_i', \underset{\sim}{v}_i', (\text{vech } \underset{\sim}{\Sigma}_i)', M_i)$, $i = 1,2,\ldots, N_n$, are a random sample from an infinite population with finite fourth moments. Assume that there is a positive real number K such that $M_i \leq K$, for every i .

Define, for $k = 1,2,\ldots, c$,

$$\underset{\sim}{\mu}_k = [E(\sum_{j=1}^{M_i} Y_{kij})]^{-1}\{E(\sum_{j=1}^{M_i} Y_{kij}\underset{\sim}{X}_{ij})\} ,$$

$$\underset{\sim}{\Sigma}_k = [E(\sum_{j=1}^{M_i} Y_{kij})]^{-1}\{E[\sum_{j=1}^{M_i} Y_{kij}(\underset{\sim}{X}_{ij} - \underset{\sim}{\mu}_k)(\underset{\sim}{X}_{ij} - \underset{\sim}{\mu}_k)']\} ,$$

$$f_k = [E(M_i)]^{-1} E(\sum_{j=1}^{M_i} Y_{kij}) .$$

Let $\hat{\underset{\sim}{\mu}}_k$, $\hat{\underset{\sim}{\Sigma}}_k$, f_k be defined by (3.D.7) - (3.D.9). Then,

$$\hat{\underset{\sim}{\mu}}_k = \underset{\sim}{\mu}_k + O_p(n^{-1/2}) ,$$

$$\hat{\underset{\sim}{\Sigma}}_k = \underset{\sim}{\Sigma}_k + O_p(n^{-1/2}) ,$$

$$\hat{f}_k = f_k + O_p(n^{-1/2}) .$$

Proof. For notational convenience, we drop the subscript n from N_n in subsequent discussion. Now,

$$\hat{f}_k - f_k = [\sum_{i=1}^{n} M_i]^{-1} \sum_{i=1}^{n} (Y_{ki.} - M_i f_k),$$

where

$$Y_{ki.} = \sum_{j=1}^{M_i} Y_{kij} \text{ , for } i = 1,2,\ldots, n .$$

The quantities, $Y_{ki.}$, $i = 1,2,\ldots, n$, are independently and identically distributed with mean $E(Y_{ki.})$ and finite variance $E(Y_{ki.}^2) - [E(Y_{ki.})]^2$. Also, $E(Y_{ki.}) = E(M_i)f_k$. Therefore, by the Central Limit Theorem,

$$n^{-1/2} \sum_{i=1}^{n} (Y_{ki.} - M_i f_k) \xrightarrow{L} N(0, \{E(Y_{ki.}^2) - [E(Y_{ki.})]^2\}) .$$

Therefore,

$$\hat{f}_k - f_k = O_p(n^{-1/2}) ,$$

since

$$n^{-1} \sum_{i=1}^{n} M_i \xrightarrow{p} E(M_i) > 0 .$$

Also, we are led to

$$n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij} \xrightarrow{p} E(Y_{ki\cdot}) > 0 ,$$

which implies

$$(n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij})^{-1} = O_p(1) .$$

Following the same arguments, we have

$$n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij}(\underset{\sim}{X}_{ij} - \mu_k) = O_p(n^{-1/2}) ,$$

which implies

$$\hat{\mu}_k - \mu_k = O_p(n^{-1/2}) .$$

Now,

$$\hat{\underset{\sim}{\Sigma}}_k - \underset{\sim}{\Sigma}_k = [\sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij} - 1]^{-1} \{ \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij}[(\underset{\sim}{X}_{ij} - \hat{\mu}_k)(\underset{\sim}{X}_{ij} - \hat{\mu}_k)' - \underset{\sim}{\Sigma}_k] + \underset{\sim}{\Sigma}_k \}$$

$$= [n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij} - n^{-1}]\{n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij}[(\underset{\sim}{X}_{ij}$$

$$- \hat{\underset{\sim}{\mu}}_k)(\underset{\sim}{X}_{ij} - \hat{\underset{\sim}{\mu}}_k)' - \underset{\sim}{\Sigma}_k]\} + o_p(n^{-1})$$

$$= [n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij} - n^{-1}]^{-1}\{n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij}[(\underset{\sim}{X}_{ij}$$

$$- \underset{\sim}{\mu}_k)(\underset{\sim}{X}_{ij} - \underset{\sim}{\mu}_k)' - \underset{\sim}{\Sigma}_k]\} + o_p(n^{-1}) ,$$

since,

$$[n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} Y_{kij}](\hat{\underset{\sim}{\mu}}_k - \underset{\sim}{\mu}_k)(\hat{\underset{\sim}{\mu}}_k - \underset{\sim}{\mu}_k)' = o_p(n^{-1}) .$$

Now,

$$E\{\sum_{j=1}^{M_i} Y_{kij}[(\underset{\sim}{X}_{ij} - \underset{\sim}{\mu}_k)(\underset{\sim}{X}_{ij} - \underset{\sim}{\mu}_k)' - \underset{\sim}{\Sigma}_k]\} = 0 .$$

Let

$$\underset{\sim}{T}_{ij} = (\underset{\sim}{X}_{ij} - \underset{\sim}{\mu}_k)(\underset{\sim}{X}_{ij} - \underset{\sim}{\mu}_k)' - \underset{\sim}{\Sigma}_k ,$$

$$\underset{\sim}{A}_{ij\ell} = (\text{vech } \underset{\sim}{T}_{ij})(\text{vech } \underset{\sim}{T}_{i\ell})' .$$

Also, let $a_{ij\ell(st)}$ be the st-th element of $\underset{\sim}{A}_{ij\ell}$. Then, there exists a positive real number K_1 such that

$$E\{\sum_{j=1}^{M_i} \sum_{\ell=1}^{M_i} Y_{kij}Y_{ki\ell}a_{ij\ell(st)}\}$$

$$\leq E\{\sum_{j=1}^{M_i} \sum_{\ell=1}^{M_i} |a_{ij\ell(st)}|\}$$

$$\leq K_1 ,$$

which implies

$$E[\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij}(\text{vech } \underset{\sim}{T}_{ij})\}\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij}(\text{vech } \underset{\sim}{T}_{ij})'\}]$$

$$= O(n^{-1}) .$$

Thus,

$$\hat{\underset{\sim}{\Sigma}}_k - \underset{\sim}{\Sigma}_k = O_p(n^{-1/2}) ,$$

since

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{kij} = O_p(1) . \qquad \square$$

By Theorem 3.D.3, the errors of the estimators $(\hat{\underset{\sim}{\mu}}_k, \hat{\underset{\sim}{\Sigma}}_k, \hat{f}_k)$ given in (3.D.7) - (3.D.9) are of order $n^{-1/2}$. In the following section, we shall see that this order of magnitude plays an important role in constructing a "good" regression estimator by using estimated auxiliary variables. In other words, the size of the effect due to the estimation of auxiliary variables will depend on the order of the error in estimating the parameters of the function used to construct the variables. In Section III.E, we give some general results for regression estimation with estimated auxiliary variables. We return to the conditional probability auxiliary variable in Section III.F.

E. The Regression Estimator with Auxiliary Variates Estimated

In this section, we shall derive some properties of the regression estimator that uses estimated auxiliary variates. For simplicity, let us consider the situation under which a simple random sample of n units is drawn without replacement from a population. This population is a set of individuals indexed by the subscript t, where $t = 1,2,\ldots, N$. Without loss of generality, the sample units are assumed to be the first n individuals. Let us denote by Y the variable of interest. Therefore, $(Y_1,\ldots, Y_n)$ is available for n sampled individuals. We also assume that a p-dimensional column vector $\underset{\sim}{X}$ of auxiliary information is available for the entire population. In other words, the value of $\underset{\sim}{X}$ is known for each population unit. We believe that some q-dimensional function $g(\underset{\sim}{X};\chi)$ of $\underset{\sim}{X}$ indexed by the unknown d-dimensional parameter vector χ, where χ is in a parameter space Γ, will be appropriate in constructing the regression estimator of the population mean $\bar{Y}_N$. The form of $g(\underset{\sim}{X};\chi)$ is assumed to be known and the function g has continuous first and second derivatives with respect to χ for all $\chi \in \Gamma$ and $\underset{\sim}{X}$ in some space χ. Let $\chi = (\gamma_1,\ldots, \gamma_d)'$ and let the true value of χ be $\chi_0 = (\gamma_{10}, \gamma_{20},\ldots, \gamma_{d0})'$. The unknown parameter vector χ_0 can be estimated by $\hat{\chi}$, which is a function of $\underset{\sim}{Z}_t = (Y_t, \underset{\sim}{X}_t')'$. For each t, let

$$\underset{\sim}{U}_t = g(\underset{\sim}{X}_t;\gamma_0) = [g_1(\underset{\sim}{X}_t;\gamma_0),\ldots, g_q(\underset{\sim}{X}_t;\gamma_0)]' ,$$

$$\hat{\underset{\sim}{U}}_t = g(\underset{\sim}{X}_t;\hat{\gamma}) ,$$

and

$$\underset{\sim}{C}'_{it} = \left(\frac{\partial\, g_i(\underset{\sim}{X}_t;\gamma_0)}{\partial\, \gamma_1} ,\ldots, \frac{\partial\, g_i(\underset{\sim}{X}_t;\gamma_0)}{\partial\, \gamma_d} \right), \qquad (3.E.1)$$

where

$\frac{\partial\, g_i(\underset{\sim}{X}_t;\gamma_0)}{\partial\, \gamma_j}$ is the value of the partial derivative $\frac{\partial\, g_i(\underset{\sim}{X}_t;\gamma)}{\partial\, \gamma_j}$

evaluated at $\gamma_j = \gamma_{j0}$, $j = 1,\ldots, d$; $i = 1,2,\ldots, q$.

Then a linear regression estimator of $\bar{Y}_N$ constructed by using $g(\underset{\sim}{X};\hat{\gamma})$ as the auxiliary vector is

$$\hat{\bar{Y}}_N = \bar{Y}_n + \underset{\sim}{B}'(\hat{\bar{\underset{\sim}{U}}}_N - \hat{\bar{\underset{\sim}{U}}}_n) ,$$

where $\underset{\sim}{B}'$ is either a preassigned constant vector or is estimated from the sample, and

$$\hat{\bar{\underset{\sim}{U}}}_N = N^{-1} \sum_{t=1}^{N} \hat{\underset{\sim}{U}}_t , \qquad (3.E.2)$$

$$\hat{\underset{\sim}{U}}_n = n^{-1} \sum_{t=1}^{n} \hat{\underset{\sim}{U}}_t \ . \qquad (3.E.3)$$

As the estimator of the population total Y_T, we take $\hat{Y}_T = N \hat{\bar{Y}}_N$.

With the auxiliary variates estimated, complexities are introduced so that the known theoretical results on regression estimation are not immediately applicable.

In this section, we investigate the limiting behavior of estimators as both the sample size and the population size become large. We assume the finite population is a random sample from an infinite population. We assume:

(i) $g(\underset{\sim}{X};\underset{\sim}{\gamma})$ is continuous and has continuous first and second derivatives with respect to $\underset{\sim}{\gamma}$ for all $\underset{\sim}{\gamma} \in \Gamma$ and $\underset{\sim}{X} \in \chi$. Let $\underset{\sim}{U} = g(\underset{\sim}{X}; \underset{\sim}{\gamma}_0)$.

(ii) Let $\{\xi_n : n = 1,2,\ldots,\}$ be a sequence of finite populations, where ξ_n is a random sample of size N_n, $N_n > N_{n-1}$, selected from a multivariate infinite population. Let the infinite population be such that the random vector $[Y, \underset{\sim}{U}', \nabla g(\underset{\sim}{X}; \underset{\sim}{\gamma}_0)]$ has finite fourth moments, where $\nabla g(\underset{\sim}{X}; \underset{\sim}{\gamma}_0)$ is the random row vector of the first partial derivatives of $g(\underset{\sim}{X}; \underset{\sim}{\gamma})$ with respect to $\underset{\sim}{\gamma}$, evaluated at $\underset{\sim}{\gamma} = \underset{\sim}{\gamma}_0$.

(iii) $\hat{\underset{\sim}{\gamma}} = \underset{\sim}{\gamma}_0 + O_p(n^{-1/2})$.

We proved in Theorem 3.D.3 that assumption (iii) holds for the parameters of the conditional probability function. In subsequent discussion, we simplify the notation by dropping the subscript n from N_n.

Theorem 3.E.1. Let the assumptions (i), (ii), and (iii) hold. Let $\underset{\sim}{B}$ be a preassigned constant vector. Assume a simple random sample of size n is chosen from N_n. Let the linear regression estimator of $\bar{Y}_N$ be

$$\hat{\bar{Y}}_N = \bar{Y}_n + \underset{\sim}{B}'(\hat{\bar{\underset{\sim}{U}}}_N - \hat{\bar{\underset{\sim}{U}}}_n)$$

where $\hat{\bar{\underset{\sim}{U}}}_N$ is defined in (3.E.2) and $\hat{\bar{\underset{\sim}{U}}}_n$ is defined in (3.E.3). Then,

$$\hat{\bar{Y}}_N - \bar{Y}_N = \tilde{\bar{Y}}_N - \bar{Y}_N + o_p(n^{-1/2}) ,$$

where $\tilde{\bar{Y}}_N$ is the usual regression estimator of $\bar{Y}_N$ for the case where the auxiliary variable $\underset{\sim}{U}$ is known; namely,

$$\tilde{\bar{Y}}_N = \bar{Y}_n + \underset{\sim}{B}'(\bar{\underset{\sim}{U}}_N - \bar{\underset{\sim}{U}}_n) .$$

Furthermore, as $n, N \to \infty$, and $\lim_{n\to\infty} (n N^{-1}) = h$, where $0 \leq h < 1$,

$$n^{1/2}(\hat{\bar{Y}}_N - \bar{Y}_N) \xrightarrow{L} N[0, (1-h)G] ,$$

where

$$G = E\{[Y_t - \underset{\sim}{B}'\underset{\sim}{U}_t]^2\} - [E\{Y_t - \underset{\sim}{B}'\underset{\sim}{U}_t\}]^2 .$$

Proof. The error of $\hat{\bar{Y}}_N$ is

$$\hat{\bar{Y}}_N - \bar{Y}_N = \bar{Y}_n - \bar{Y}_N - \underset{\sim}{B}'(\hat{\bar{\underset{\sim}{U}}}_n - \hat{\bar{\underset{\sim}{U}}}_N)$$

$$= \bar{Y}_n - \bar{Y}_N - \underset{\sim}{B}'(\bar{\underset{\sim}{U}}_n - \bar{\underset{\sim}{U}}_N) - \underset{\sim}{B}'(\hat{\bar{\underset{\sim}{U}}}_n - \bar{\underset{\sim}{U}}_n - \hat{\bar{\underset{\sim}{U}}}_N + \bar{\underset{\sim}{U}}_N) .$$

Since $g(\underset{\sim}{X};\gamma)$ has continuous first and second derivatives at $\gamma = \gamma_0$, Taylor's formula gives

$$\hat{\bar{\underset{\sim}{U}}}_n = \bar{\underset{\sim}{U}}_n + \bar{\underset{\sim}{C}}_n(\hat{\gamma} - \gamma_0) + o_p(n^{-1/2}) , \qquad (3.E.4)$$

where $\underset{\sim}{C}_{it}$ is defined in (3.E.1),

$$\underset{\sim}{C}_t = (\underset{\sim}{C}_{1t}, \ldots, \underset{\sim}{C}_{qt})' ,$$

and

$$\bar{\underset{\sim}{C}}_n = n^{-1} \sum_{t=1}^{n} \underset{\sim}{C}_t .$$

Similarly,

$$\hat{\bar{U}}_N = \bar{U}_N + \bar{C}_N(\hat{\gamma} - \gamma_0) + o_p(N^{-1/2}) , \qquad (3.E.5)$$

where

$$\bar{C}_N = N^{-1} \sum_{t=1}^{N} C_t .$$

Combining (3.E.4) and (3.E.5), we have

$$\hat{\bar{U}}_n - \bar{U}_n - \bar{U}_N + \hat{\bar{U}}_N = (\bar{C}_n - \bar{C}_N)(\hat{\gamma} - \gamma_0) + o_p(n^{-1/2}) .$$

To establish the order of magnitude for $\hat{\bar{U}}_n - \bar{U}_n - \bar{U}_N + \hat{\bar{U}}_N$, we consider the first and second moments of $\bar{C}_n - \bar{C}_N$. In simple random sampling,

$$E(\bar{C}_n - \bar{C}_N) = E\{E(\bar{C}_n - \bar{C}_N) \mid \xi_n\}$$

$$= 0 ,$$

where $E\{. \mid \xi_n\}$ means that the expectation is with respect to simple random sampling from the fixed finite population ξ_n. Also,

$$E[(\bar{C}_n - \bar{C}_N)(\bar{C}_n - \bar{C}_N)'] = E\{E[(\bar{C}_n - \bar{C}_N)(\bar{C}_n - \bar{C}_N)' \mid \xi_n]\}$$

$$= n^{-1}(1 - N^{-1}n)\, E[(N-1)^{-1} \sum_{t=1}^{N} (C_t$$

$$- \bar{\underset{\sim}{C}}_N)(\underset{\sim}{C}_t - \bar{\underset{\sim}{C}}_N)']$$

$$= O(n^{-1}) ,$$

since the assumption (ii) implies that

$$E[(N-1)^{-1} \sum_{t=1}^{N} (\underset{\sim}{C}_t - \bar{\underset{\sim}{C}}_N)(\underset{\sim}{C}_t - \bar{\underset{\sim}{C}}_N)']$$

is $O(1)$. By using Corollary 5.1.1.1 in Fuller (1976),

$$\bar{\underset{\sim}{C}}_n - \bar{\underset{\sim}{C}}_N = o_p(n^{-1/2})$$

and

$$n^{1/2}(\hat{\bar{Y}}_N - \bar{Y}_N) = n^{1/2}(\tilde{\bar{Y}}_N - \bar{Y}_N) + o_p(1) .$$

This implies that $n^{1/2}(\hat{\bar{Y}}_N - \bar{Y}_N)$ has the same limiting distribution as $n^{1/2}(\tilde{\bar{Y}}_N - \bar{Y}_N)$ provided that the limiting distribution of the latter exists. Let $e_i = Y_i - \underset{\sim}{B}'\underset{\sim}{U}_i - E(Y_i) - \underset{\sim}{B}'E(\underset{\sim}{U}_i)$, $i = 1,2,\ldots, N$. Then,

$$n^{1/2}(\tilde{\bar{Y}}_N - \bar{Y}_N) = n^{-1/2}(1 - N^{-1} n) \sum_{i=1}^{n} e_i$$

$$- (N^{-1} n)^{1/2}(1 - N^{-1} n)^{1/2}(N - n)^{-1/2} \sum_{i=n+1}^{N} e_i .$$

Since $[Y_1, \underset{\sim}{U}_1'], \ldots, [Y_N, \underset{\sim}{U}_N']$ are independently and identically distributed with finite mean and finite second moment, it follows from Multivariate Central Limit Theorem that

$$S_{1n} = n^{-1/2}(1 - N^{-1}n) \sum_{i=1}^{n} e_i$$

converges in distribution to a normal random variable with mean zero and variance $(1 - h)^2 G$. In a similar manner,

$$S_{2n} = (N^{-1}n)^{1/2}(1 - N^{-1}n)^{1/2}(N - n)^{-1/2} \sum_{i=n+1}^{N} e_i$$

converges in distribution to a normal random variable with mean zero and variance $h(1 - h)G$. Since S_{1n} and S_{2n} are independent, the result follows. □

In most applications, $\underset{\sim}{B}$ is estimated from the sample. For instance, an effective estimator is the familiar least squares estimator of $\underset{\sim}{B}$.

Theorem 3.E.2. Let the assumptions (i), (ii), and (iii) hold. Let the estimator $\hat{\underset{\sim}{B}}$ satisfy $\hat{\underset{\sim}{B}} - \underset{\sim}{B} = O_p(n^{-1/2})$. Then, in simple random sampling, the linear regression estimator of $\bar{Y}_N$, given by

$$\hat{\bar{Y}}_{N,\ell} = \bar{Y}_n + \hat{\underset{\sim}{B}}'(\hat{\bar{\underset{\sim}{U}}}_N - \hat{\bar{\underset{\sim}{U}}}_n)$$

satisfies

$$\hat{\bar{Y}}_{N,\ell} - \bar{Y}_N = \tilde{\bar{Y}}_N - \bar{Y}_N + o_p(n^{-1/2}) ,$$

where $\tilde{\bar{Y}}_N$ is defined as in Theorem 3.E.1. Furthermore, as $n, N \to \infty$, with $\lim_{n\to\infty} (n N^{-1}) = h$, where $0 \leq h < 1$,

$$n^{1/2} (\hat{\bar{Y}}_{N,\ell} - \bar{Y}_N) \xrightarrow{L} N[0, (1-h)G] ,$$

where G is defined in Theorem 3.E.1.

Proof. The error of $\hat{\bar{Y}}_{N,\ell}$ can be expressed as

$$\hat{\bar{Y}}_{N,\ell} - \bar{Y}_N = \bar{Y}_n - \bar{Y}_N + \underset{\sim}{B}'(\bar{\underset{\sim}{U}}_N - \bar{\underset{\sim}{U}}_n) + \underset{\sim}{B}'(\hat{\bar{\underset{\sim}{U}}}_N - \bar{\underset{\sim}{U}}_N - \hat{\bar{\underset{\sim}{U}}}_n + \bar{\underset{\sim}{U}}_n)$$

$$+ (\hat{\underset{\sim}{B}} - \underset{\sim}{B})'(\hat{\bar{\underset{\sim}{U}}}_N - \bar{\underset{\sim}{U}}_N - \hat{\bar{\underset{\sim}{U}}}_n + \bar{\underset{\sim}{U}}_n)$$

$$+ (\hat{\underset{\sim}{B}} - \underset{\sim}{B})'(\bar{\underset{\sim}{U}}_N - \bar{\underset{\sim}{U}}_n) .$$

As shown in Theorem 3.E.1, $\hat{\bar{\underset{\sim}{U}}}_N - \bar{\underset{\sim}{U}}_N - \hat{\bar{\underset{\sim}{U}}}_n + \bar{\underset{\sim}{U}}_n = O_p(n^{-1}) + o_p(n^{-1/2})$. Also, in simple random sampling,

$$\bar{\underset{\sim}{U}}_n - \bar{\underset{\sim}{U}}_N = O_p(n^{-1/2}) .$$

Therefore,

$$\hat{\bar{Y}}_{N,\ell} - \bar{Y}_N = \tilde{\bar{Y}}_N - \bar{Y}_N + o_p(n^{-1}) + o_p(n^{-1/2})$$

$$= \tilde{\bar{Y}}_N - \bar{Y}_N + o_p(n^{-1/2}) \ .$$

The asymptotic normality follows from the same arguments used in Theorem 3.E.1. □

In practice, one commonly used estimator of $\underset{\sim}{B}$ is the least squares estimator $\tilde{\underset{\sim}{B}}$, where

$$\tilde{\underset{\sim}{B}} = (\sum_{t=1}^{n} \hat{\underset{\sim}{U}}_t \hat{\underset{\sim}{U}}_t')^{-1} (\sum_{t=1}^{n} \hat{\underset{\sim}{U}}_t Y_t) \ . \tag{3.E.6}$$

The following theorem investigates the effect due to the estimation of $\underset{\sim}{U}_t$ on the least squares estimator $\tilde{\underset{\sim}{B}}$.

Theorem 3.E.3. Let assumptions (i), (ii), and (iii) hold. The estimator $\tilde{\underset{\sim}{B}}$ of $\underset{\sim}{B}$ is defined in (3.E.6). Let $(\underset{\sim}{C}_t', \underset{\sim}{U}_t')'$ be defined in (3.E.1). Assume that

$$\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' = \underset{\sim}{M}_1 \ ,$$

where $\underset{\sim}{M}_1$ is positive definite. Then,

$$\tilde{\underset{\sim}{B}} - \underset{\sim}{B} = (\underset{\sim}{b} - \underset{\sim}{B}) + o_p(n^{-1/2}) \ ,$$

where $\underset{\sim}{b}$ is the usual least squares estimator of $\underset{\sim}{B}$ when each $\underset{\sim}{U}_t$ is known; i.e.,

$$\underset{\sim}{b} = \left(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' \right)^{-1} \left(\sum_{t=1}^{n} \underset{\sim}{U}_t Y_t \right) .$$

Proof. Since $g(\underset{\sim}{X};\gamma)$ has continuous second derivatives at

$$\gamma = \gamma_0 ,$$

$$\begin{aligned}
\tilde{\underset{\sim}{B}} - \underset{\sim}{B} &= \left(\sum_{t=1}^{n} \hat{\underset{\sim}{U}}_t \hat{\underset{\sim}{U}}_t' \right)^{-1} \left[\sum_{t=1}^{n} \hat{\underset{\sim}{U}}_t (Y_t - \hat{\underset{\sim}{U}}_t' \underset{\sim}{B}) \right] \\
&= \left(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' \right)^{-1} \left[\sum_{t=1}^{n} \underset{\sim}{U}_t (Y_t - \underset{\sim}{U}_t' \underset{\sim}{B}) \right] \\
&\quad + \left\{ \left(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' \right)^{-1} \left[- \sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{C}_t' \underset{\sim}{B} + \sum_{t=1}^{n} \underset{\sim}{C}_t (Y_t - \underset{\sim}{U}_t' \underset{\sim}{B}) \right] \right. \\
&\quad \left. - \left(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' \right)^{-1} \left(\sum_{t=1}^{n} \underset{\sim}{C}_t \underset{\sim}{U}_t' + \sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{C}_t' \right) \left(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' \right)^{-1} \right. \\
&\quad \left. \times \left[\sum_{t=1}^{n} \underset{\sim}{U}_t (Y_t - \underset{\sim}{U}_t \underset{\sim}{B}') \right] \right\} (\hat{\gamma} - \gamma_0) \\
&\quad + o_p(n^{-1/2}) ,
\end{aligned}$$

using Taylor's approximation. Now,

$$\left[n^{-1} \sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' \right]^{-1} = O_p(1)$$

because

$$\text{plim } n^{-1} \sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t' = \underset{\sim}{M}_1$$

which is positive definite. Then,

$$\begin{aligned}
&\Big(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t'\Big)^{-1}\Big[-\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{C}_t' \underset{\sim}{B} + \sum_{t=1}^{n} \underset{\sim}{C}_t (Y_t - \underset{\sim}{U}_t' \underset{\sim}{B})\Big] \\
&\qquad - \Big(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t'\Big)^{-1}\Big(\sum_{t=1}^{n} \underset{\sim}{C}_t \underset{\sim}{U}_t' + \sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{C}_t'\Big)\Big(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t'\Big)^{-1} \\
&\qquad \times \Big[\sum_{t=1}^{n} \underset{\sim}{U}_t (Y_t - \underset{\sim}{U}_t' \underset{\sim}{B})\Big] \\
&= \Big(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t'\Big)^{-1} \sum_{t=1}^{n} \underset{\sim}{C}_t Y_t - \Big(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t'\Big)^{-1}\Big(\sum_{t=1}^{n} \underset{\sim}{C}_t \underset{\sim}{U}_t' + \sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{C}_t'\Big) \\
&\qquad \times \Big(\sum_{t=1}^{n} \underset{\sim}{U}_t \underset{\sim}{U}_t'\Big)^{-1} \sum_{t=1}^{n} \underset{\sim}{U}_t Y_t \\
&= o_p(1) ,
\end{aligned}$$

since assumption (ii) implies that

$$n^{-1} \sum_{t=1}^{n} \underset{\sim}{C}_t Y_t = o_p(1) ,$$

$$n^{-1} \sum_{t=1}^{n} \underset{\sim}{U}_t Y_t = o_p(1) ,$$

and

$$n^{-1} \sum_{t=1}^{n} \underset{\sim}{C}_t \underset{\sim}{U}_t' = O_p(1) .$$

Therefore,

$$\tilde{\underset{\sim}{B}} - \underset{\sim}{B} = \underset{\sim}{b} - \underset{\sim}{B} + O_p(n^{-1/2}) . \qquad \square$$

In summary, the above three theorems make the following points:

(1) In estimating the population mean $\bar{Y}_N$, the construction of $\hat{\underset{\sim}{U}}$ affects the regression estimator of $\bar{Y}_N$ by a term of order in probability n^{-1} when the regression coefficient $\underset{\sim}{B}$ is either a preassigned constant vector or is estimated with error of order $n^{-1/2}$. For the large sample case, the bias is negligible, and the large sample variance is equal to the large sample variance of the regression estimator with known auxiliary variables.

(2) In estimating the regression coefficient vector $\underset{\sim}{B}$, the estimation of $\underset{\sim}{U}$ will affect the least squares estimator of $\underset{\sim}{B}$ by a term of order $n^{-1/2}$. The order of the effect is the same as that of the error of the least squares estimator $\underset{\sim}{b}$ for the case where the auxiliary variable $\underset{\sim}{U}$ is known.

F. Regression Estimation with Estimated Conditional Probability as the Auxiliary Variable

Now, let us return to the original problem stated in Section III.A, where a sample of n clusters is drawn from the population of N clusters by the method of simple random sampling. Although $\underset{\sim}{X}$ is available for each element of the N clusters, we believe that the posterior probability transformation of $\underset{\sim}{X}$ is more appropriate than $\underset{\sim}{X}$ itself as an auxiliary variable in constructing the regression estimator of the population Y-total for each category. For each category, the posterior probability transformation of $\underset{\sim}{X}$ is defined to be

$$Z_k = \Big[\sum_{i=1}^{c} f_i |\underset{\sim}{\Sigma}_i|^{-1/2} \exp[-1/2\,(\underset{\sim}{X} - \underset{\sim}{\mu}_i)'\underset{\sim}{\Sigma}_i^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}_i)]\Big]^{-1}$$

$$\times f_k |\underset{\sim}{\Sigma}_k|^{-1/2} \exp[-1/2\,(\underset{\sim}{X} - \underset{\sim}{\mu}_k)'\underset{\sim}{\Sigma}_k^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}_k)] \quad , \quad k = 1,2,\ldots, c. \tag{3.F.1}$$

The posterior probability transformation is estimated by estimating the parameters $\underset{\sim}{\mu}_k$, $\underset{\sim}{\Sigma}_k$, f_k , $k = 1,2,\ldots, c$, using estimators $\hat{\underset{\sim}{\mu}}_k$, $\hat{\underset{\sim}{\Sigma}}_k$ and $\hat{f}_k$ defined in (3.D.7), (3.D.8), and (3.D.9), respectively. The sum of the estimated posterior probabilities for each cluster serves as the auxiliary variable. In investigating the large sample properties of the regression estimators, we specify a sequence of finite populations, which are random samples from some infinite population.

Recall that for $k = 1,2,\ldots, c$, the variables $\underset{\sim}{X}_{ij}$, Y_{kij}, Z_{kij}, $Y_{ki.}$, $Z_{ki.}$, $\bar{Y}_{k..}^{(n)}$, $\bar{Z}_{k..}^{(N)}$, $\bar{\bar{Z}}_{k..}^{(N)}$, are defined in Section III.A and Section III.B. We demonstrate the asymptotic properties of the regression estimator $\tilde{\bar{\bar{Y}}}_{k..(\ell r)}$ defined in (3.B.3) in the following theorem.

Theorem 3.F.1. Let $\{\xi_n: n = 1,2,\ldots,\}$ be a sequence of finite populations, where ξ_n contains N_n clusters, $N_n > N_{n-1}$. Let M_i denote the number of secondary elements in cluster i for each i. A sample of n clusters is selected from population ξ_n by simple random sampling without replacement. Let

$$\lim_{n\to\infty} n N_n^{-1} = h ,$$

where $0 \leq h < 1$. Let $\underset{\sim}{v}_i = (v_{1i}, v_{2i},\ldots, v_{ci})'$; let $\underset{\sim}{w}_i = (w_{1i},\ldots, w_{ci})'$; let $\underset{\sim}{\eta}_{ij} = (\eta_{1ij},\ldots, \eta_{cij})'$; let $\underset{\sim}{\alpha}_{ij} = (\alpha_{1ij},\ldots, \alpha_{cij})'$. Let

$$\underset{\sim}{X}_{ij} = \underset{\sim}{u}_i + \underset{\sim}{\varepsilon}_{ij} ,$$

$$Y_{kij} = v_{ki} + \eta_{kij} ,$$

$$Z_{kij} = w_{ki} + \alpha_{kij} \ , \quad j = 1,2,\ldots, M_i \ ; \quad i = 1,2,\ldots, N_n \ ,$$

$$k = 1,2,\ldots, c \ ,$$

where the random vector $(\underset{\sim}{u}_i', \underset{\sim}{v}_i', \underset{\sim}{w}_i')$ is the 'primary component' and the random vector $(\underset{\sim}{\varepsilon}_{ij}', \underset{\sim}{\eta}_{ij}', \underset{\sim}{\alpha}_{ij}')$ is the 'secondary component'. We assume:

(i) For cluster i, the $(\underset{\sim}{\varepsilon}_{ij}', \underset{\sim}{\eta}_{ij}', \underset{\sim}{\alpha}_{ij}')'$, $j = 1,2,\ldots, M_i$, are a random sample from an infinite population with zero mean vector, uniformly bounded fourth moments, and positive definite covariance matrix $\underset{\sim}{\Sigma}_i$, where

$$\underset{\sim}{\Sigma}_i = \begin{pmatrix} \underset{\sim}{\Sigma}_{\varepsilon\varepsilon i} & \underset{\sim}{\Sigma}_{\varepsilon\eta i} & \underset{\sim}{\Sigma}_{\varepsilon\alpha i} \\ \underset{\sim}{\Sigma}_{\varepsilon\eta i}' & \underset{\sim}{\Sigma}_{\eta\eta i} & \underset{\sim}{\Sigma}_{\eta\alpha i} \\ \underset{\sim}{\Sigma}_{\varepsilon\alpha i}' & \underset{\sim}{\Sigma}_{\eta\alpha i} & \underset{\sim}{\Sigma}_{\alpha\alpha i} \end{pmatrix} ,$$

(ii) The vectors $(\underset{\sim}{u}_i', \underset{\sim}{v}_i', \underset{\sim}{w}_i', (\text{vech } \underset{\sim}{\Sigma}_i)', M_i)$, $i = 1,2,\ldots, N_n$, are a simple random sample from an infinite population with finite fourth moments. Assume that there exists a positive real number K such that $M_i \leq K$, for every i.

Then,

$$n^{1/2}(\tilde{\bar{Y}}_{k..(\ell r)} - \bar{Y}_{k..}^{(N)}) \xrightarrow{L} N[0, (1 - h)\sigma_{yy}(1 - \rho_{yz}^2)] \ ,$$

where $\tilde{\bar{Y}}_{k..(\ell r)}$ is defined in (3.B.3),

$$\rho_{yz} = [\sigma_{yy}\sigma_{zz}]^{-1/2}\sigma_{yz} ,$$

$$\sigma_{yy} = E\{M_i^2 v_{ki}^2 + M_i\sigma_{\eta\eta kki}\} - [E\{M_i v_{ki}\}]^2 ,$$

$$\sigma_{zz} = E\{M_i^2 w_{ki}^2 + M_i\sigma_{\alpha\alpha kki}\} - [E\{M_i w_{ki}\}]^2 ,$$

$$\sigma_{yz} = E\{M_i^2 v_{ki}w_{ki} + M_i\sigma_{\eta\alpha kki}\} - [E\{M_i w_{ki}\}][E\{M_i v_{ki}\}] ,$$

$\sigma_{\eta\eta kki}$ = the kk-th element of $\underset{\sim}{\Sigma}_{\eta\eta i}$,

$\sigma_{\alpha\alpha kki}$ = the kk-th element of $\underset{\sim}{\Sigma}_{\alpha\alpha i}$,

$\sigma_{\eta\alpha kki}$ = the kk-th element of $\underset{\sim}{\Sigma}_{\eta\alpha i}$.

Proof. Let us simplify the notation in subsequent discussion by dropping the subscript n from N_n. Let

$$Y_{ki.} = \sum_{j=1}^{M_i} Y_{kij} ,$$

$$Z_{ki.} = \sum_{j=1}^{M_i} p(\theta_{ij}=k \,|\, \underset{\sim}{X}_{ij}) ,$$

$$W_{ki.} = \sum_{j=1}^{M_i} \hat{p}(\theta_{ij}=k \,|\, \underset{\sim}{X}_{ij}) ,$$

$$\bar{Y}_{k..}^{(n)} = n^{-1} \sum_{i=1}^{n} Y_{ki.} ,$$

$$\bar{Y}_{k..}^{(N)} = N^{-1} \sum_{i=1}^{N} Y_{ki.} ,$$

$$\bar{Z}_{k..}^{(n)} = n^{-1} \sum_{i=1}^{n} Z_{ki.} ,$$

$$\bar{Z}_{k..}^{(N)} = N^{-1} \sum_{i=1}^{N} Z_{ki.} ,$$

$$\bar{W}_{k..}^{(n)} = n^{-1} \sum_{i=1}^{n} W_{ki.} ,$$

and

$$\bar{W}_{k..}^{(N)} = N^{-1} \sum_{i=1}^{N} W_{ki.} .$$

Since the posterior probability defined in (3.F.1) has continuous derivatives of all orders with respect to χ_0 ,

$$\bar{W}_{k..}^{(n)} = \bar{Z}_{k..}^{(n)} + \bar{D}_{\sim k..}^{(n)}(\hat{\chi} - \chi_0) + O_p(n^{-1})$$

and

$$\bar{W}_{k..}^{(N)} = \bar{Z}_{k..}^{(N)} + \bar{D}_{\sim k..}^{(N)}(\hat{\chi} - \chi_0) + O_p(N^{-1}) ,$$

where

$$\chi_0 = (\mu_1', \ldots, \mu_c', (\text{vech } \Sigma_{\sim 1})', \ldots, (\text{vech } \Sigma_{\sim c})', f_1, \ldots, f_c)' ,$$

$$\hat{\chi} = (\hat{\underset{\sim}{\mu}}_1', \ldots, \hat{\underset{\sim}{\mu}}_c', (\text{vech } \hat{\underset{\sim}{\Sigma}}_1)', \ldots, (\text{vech } \hat{\underset{\sim}{\Sigma}}_c)', \hat{f}_1, \ldots, \hat{f}_c)' ,$$

$$\bar{\underset{\sim}{D}}_{k..}^{(n)} = n^{-1} \sum_{i=1}^{n} \underset{\sim}{D}_{ki.} ,$$

$$\bar{\underset{\sim}{D}}_{k..}^{(N)} = N^{-1} \sum_{i=1}^{N} \underset{\sim}{D}_{ki.} ,$$

$$\underset{\sim}{D}_{ki.} = \sum_{j=1}^{M_i} \left[\left(\frac{\partial\, p(\theta=k|\underset{\sim}{X}_{ij})}{\partial\, \underset{\sim}{\mu}_1} \right)', \ldots\ldots, \left(\frac{\partial\, p(\theta=k|\underset{\sim}{X}_{ij})}{\partial\, \underset{\sim}{\mu}_c} \right)' , \right.$$

$$\left(\frac{\partial\, p(\theta=k|\underset{\sim}{X}_{ij})}{\partial\, \text{vech } \underset{\sim}{\Sigma}_1} \right)', \ldots\ldots, \left(\frac{\partial\, p(\theta=k|\underset{\sim}{X}_{ij})}{\partial\, \text{vech } \underset{\sim}{\Sigma}_c} \right)' ,$$

$$\left. \frac{\partial\, p(\theta=k|\underset{\sim}{X}_{ij})}{\partial\, f_1}, \ldots\ldots, \frac{\partial\, p(\theta=k|\underset{\sim}{X}_{ij})}{\partial\, f_c} \right] ,$$

$i = 1,2,\ldots, N$, $k = 1,2,\ldots, c$. Let $\underset{\sim}{\lambda}_1 = (\lambda_{11}, \ldots, \lambda_{1p})'$ be an arbitrary constant vector, and let $\underset{\sim}{\lambda}_2 = (\lambda_{21}, \lambda_{22}, \lambda_{23}, \ldots, \lambda_{2s})'$ be an arbitrary constant vector, where $s = p(p+1)/2$. Then, from (3.D.1) - (3.D.7), for each $\ell(\ell=1,2,\ldots, c)$,

$$\left| \underset{\sim}{\lambda}_1' \frac{\partial\, p(\theta=k|\underset{\sim}{X})}{\partial\, \underset{\sim}{\mu}_\ell} \right| \leq \left| \underset{\sim}{\lambda}_1' \underset{\sim}{\Sigma}_\ell^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu}_\ell) \right| ,$$

$$\left| \underset{\sim}{\lambda}_2' \frac{\partial\, p(\theta=k|\underset{\sim}{X})}{\partial\, \text{vech}\, \underset{\sim}{\Sigma}_\ell} \right| \leq \left| \underset{\sim}{\lambda}_2' \{ \underset{\sim}{\Phi}' (\underset{\sim}{\Sigma}_\ell^{-1} \otimes \underset{\sim}{\Sigma}_\ell^{-1}) \underset{\sim}{\Phi}\, \text{vech}[(\underset{\sim}{X} - \underset{\sim}{\mu}_\ell)(\underset{\sim}{X} - \underset{\sim}{\mu}_\ell)'] \right.$$

$$\left. - 2\text{vech}\, \underset{\sim}{\Sigma}_\ell^{-1} + \text{vech}[\text{diag}(\underset{\sim}{\Sigma}_\ell^{-1})] \} \right| \quad ,$$

$$\left| \frac{\partial\, p(\theta=k|\underset{\sim}{X})}{\partial\, f_\ell} \right| \leq f_\ell^{-1} \quad ,$$

where $f_\ell > 0$, $\ell = 1,2,\ldots, c$. Note that on the right-hand side of the above inequalities, all the terms inside the absolute value function are polynomials in $\underset{\sim}{X}$ of degree at most two. Hence, assumptions (i) - (ii) imply that

$$E\{(N-1)^{-1} \sum_{i=1}^{N} (\underset{\sim}{D}_{ki.} - \bar{\underset{\sim}{D}}_{k..}^{(N)})(\underset{\sim}{D}_{ki.} - \bar{\underset{\sim}{D}}_{k..}^{(N)})'\} = O(1) \quad ,$$

because

$$E\{N^{-1} \sum_{i=1}^{N} M_i^2\} = O(1) \quad .$$

Now,

$$E(\bar{\underset{\sim}{D}}_{k..}^{(n)} - \bar{\underset{\sim}{D}}_{k..}^{(N)})$$

$$= E\{E(\bar{\underset{\sim}{D}}_{k..}^{(n)} - \bar{\underset{\sim}{D}}_{k..}^{(N)} | \xi_n)\}$$

$$= 0 \ ,$$

and

$$E[(\underset{\sim}{\bar{D}}_{k..}^{(n)} - \underset{\sim}{\bar{D}}_{k..}^{(N)})(\underset{\sim}{\bar{D}}_{k..}^{(n)} - \underset{\sim}{\bar{D}}_{k..}^{(N)})']$$

$$= E\{E[(\underset{\sim}{\bar{D}}_{k..}^{(n)} - \underset{\sim}{\bar{D}}_{k..}^{(N)})(\underset{\sim}{\bar{D}}_{k..}^{(n)} - \underset{\sim}{\bar{D}}_{k..}^{(N)})' \,|\, \xi_n]\}$$

$$= n^{-1}(1 - n\,N^{-1})E\{(N-1)^{-1} \sum_{i=1}^{N} (\underset{\sim}{D}_{ki.} - \underset{\sim}{\bar{D}}_{k..}^{(N)})(\underset{\sim}{D}_{ki.} - \underset{\sim}{\bar{D}}_{k..}^{(N)})'\}$$

$$= O(n^{-1}) \ .$$

Thus,

$$\underset{\sim}{\bar{D}}_{k..}^{(n)} - \underset{\sim}{\bar{D}}_{k..}^{(N)} = O_p(n^{-1/2}) \ .$$

Since $\hat{\chi} - \hat{\chi}_0 = O_p(n^{-1/2})$ by Theorem 3.D.3, we have

$$(\underset{\sim}{\bar{D}}_{k..}^{(n)} - \underset{\sim}{\bar{D}}_{k..}^{(N)})(\hat{\chi} - \hat{\chi}_0) = O_p(n^{-1}) \ .$$

Following the same arguments as above, we have

$$n^{-1} \sum_{i=1}^{n} (W_{ki.} - \bar{W}_{k..}^{(n)})^2 = n^{-1} \sum_{i=1}^{n} (Z_{ki.} - \bar{Z}_{k..}^{(n)})^2 + O_p(n^{-1/2})$$

and

$$n^{-1} \sum_{i=1}^{n} (W_{ki.} - \bar{W}_{k..}^{(n)})[Y_{ki.} - \bar{Y}_{k..}^{(n)}]$$

$$= n^{-1} \sum_{i=1}^{n} (Z_{ki.} - \bar{Z}_{k..}^{(n)})[Y_{ki.} - \bar{Y}_{k..}^{(n)}]$$

$$+ o_p(n^{-1/2}) .$$

Since each $Z_{ki.}$ has common mean and common variance,

$$n^{-1} \sum_{i=1}^{n} (Z_{ki.} - \bar{Z}_{k..}^{(n)})^2 \xrightarrow{p} \sigma_{zz} ,$$

where

$$\sigma_{zz} = E[M_i^2 w_{ki}^2 + M_i \sigma_{\alpha\alpha kki}] - [E(M_i w_{ki})]^2 > 0 .$$

Thus,

$$[n^{-1} \sum_{i=1}^{n} (Z_{ki.} - \bar{Z}_{k..}^{(n)})^2]^{-1} \xrightarrow{p} \sigma_{zz}^{-1} .$$

Also,

$$n^{-1} \sum_{i=1}^{n} (Z_{ki.} - \bar{Z}_{k..}^{(n)})(Y_{ki.} - \bar{Y}_{k..}^{(n)}) \xrightarrow{p} \sigma_{yz} ,$$

where

$$\sigma_{yz} = E[M_i^2 v_{ki} w_{ki} + M_i \sigma_{\eta\alpha kki}] - E(M_i v_{ki}) E(M_i w_{ki}) .$$

It follows that

$$\tilde{\beta}_k - B_k = o_p(1) ,$$

where

$$B_k = \sigma_{zz}^{-1} \sigma_{yz} ,$$

$$\tilde{\beta}_k = [\sum_{i=1}^{n} (W_{ki.} - \bar{W}_{k..}^{(n)})^2]^{-1} \sum_{i=1}^{n} (W_{ki.} - \bar{W}_{k..}^{(n)})(Y_{ki.} - \bar{Y}_{k..}^{(n)}) .$$

Since

$$n^{1/2} (\bar{Z}_{k..}^{(n)} - \bar{Z}_{k..}^{(N)}) = o_p(1) ,$$

$$(\tilde{\beta}_k - B_k) n^{1/2} (\bar{Z}_{k..}^{(n)} - \bar{Z}_{k..}^{(N)}) = o_p(1) .$$

Therefore,

$$n^{1/2} (\tilde{\bar{Y}}_{k..(\ell r)} - \bar{Y}_{k..}^{(N)}) = n^{1/2} [\bar{Y}_{k..}^{(n)} - \bar{Y}_{k..}^{(N)} - B_k (\bar{Z}_{k..}^{(n)} - \bar{Z}_{k..}^{(N)})] + o_p(1) .$$

To obtain the asymptotic distribution of $\tilde{\bar{Y}}_{k..(\ell r)} - \bar{Y}_{k..}^{(N)}$, let

$$d_{ki} = Y_{ki.} - B_k Z_{ki.} - E(Y_{ki.}) - B_k E(Z_{ki.}) , \quad i = 1,2,\ldots, N .$$

Since $d_{k1}, d_{k2},\ldots, d_{kN}$ are i.i.d. with mean zero and variance $E(d_{ki}^2)$,

$$[\sum_{i=1}^{n} E(d_{ki}^2)]^{-1/2} \sum_{i=1}^{n} d_{ki} \xrightarrow{L} N(0, 1) .$$

Similarly,

$$[\sum_{i=n+1}^{N} E(d_{ki}^2)]^{-1/2} \sum_{i=n+1}^{N} d_{ki} \xrightarrow{L} N(0, 1) .$$

Also,

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E(d_{ki}^2) = \sigma_{yy}(1 - \rho_{yz}^2) .$$

Similarly,

$$\lim_{n\to\infty} (N - n)^{-1} \sum_{i=n+1}^{N} E(d_{ki}^2) = \sigma_{yy}(1 - \rho_{yz}^2) .$$

Since $\sum_{i=1}^{n} d_{ki}$ and $\sum_{j=n+1}^{N} d_{kj}$ are independent, the result follows. □

IV. LANDSAT CROP ESTIMATION

In this chapter, we investigate the use of the estimated posterior probability as an auxiliary variable in constructing a regression estimator of crop acreages. The posterior probability of crop j is defined to be the conditional probability that the ground cover of a pixel with the satellite value $\underset{\sim}{x}$ is crop j.

A. Data and Procedures

The LANDSAT data file used in the study is for the 1979 JES segments in Northern Missouri. The file contains a simple random sample of 46 segments with 20,262 pixels. A segment is a primary sampling unit of size about one square mile. Segment 6038 is not used in the analysis because it was incorrectly located. The sample of 19,943 pixels has an image date of August 3, 1979 and has been analyzed by using various classifiers.

The pixel data are divided into eight crop groups according to the ground identification. Let θ be used to identify the crop, where $\theta = 1,2,3,4,5,6,7,8$ refer to corn, winter wheat, pasture, soybeans, woods, alfalfa, sorghum, and all other, respectively. The basic element of LANDSAT data is the vector $\underset{\sim}{X}_{ij}$ of four bands of radiometric values. Let

$$Y_{kij} = \begin{cases} 1 & \text{if JES crop code } \theta = k \text{ for j-th pixel of segment } i \\ 0 & \text{otherwise,} \end{cases}$$

X_{1ij} = Satellite band 4 variable for pixel j in segment i,

X_{2ij} = Satellite band 5 variable for pixel j in segment i,

X_{3ij} = Satellite band 6 variable for pixel j in segment i,

X_{4ij} = Satellite band 7 variable for pixel j in segment i,

$\underset{\sim}{X}_{ij} = (X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij})'$.

Let each pixel be assigned to a crop group by a classification procedure based upon $\underset{\sim}{X}_{ij}$. Let

$$W_{kij} = \begin{cases} 1 & \text{if pixel } j \text{ of segment } i \text{ is classified as crop } k \ (\theta=k) \text{ by the USDA pixel classification} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$Y_{ki.} = \sum_j Y_{kij}$ = number of pixels identified as crop k by the ground survey in segment i .

$W_{ki.} = \sum_j W_{kij}$ = number of pixels classified as crop k in segment i by the pixel classification.

For crop k , the fraction of the area, f_k , and the class-conditional probability density $p(\underset{\sim}{X}|\theta=k)$ are to be estimated for $k = 1,2,\ldots, 8$. Let $\hat{p}(\underset{\sim}{X}|\theta=k)$ and $\hat{f}_k$ be the estimators of

$p(\underset{\sim}{X}|\theta=k)$ and f_k, respectively, $k = 1,2,\ldots, 8$. Then the estimated posterior probability that a point with a satellite value $\underset{\sim}{X}$ is from crop k is

$$\hat{p}(\theta=k|\underset{\sim}{X}) = \frac{\hat{p}(\underset{\sim}{X}|\theta=k)\hat{f}_k}{\sum_{i=1}^{8} \hat{p}(\underset{\sim}{X}|\theta=i)\hat{f}_i}, \quad k = 1,2,\ldots, 8. \tag{4.A.1}$$

To use these conditional probabilities in the construction of a regression estimator of crop acres, the sum of the conditional probabilities is created for each segment. The ground acreage of the crop in the segment is then regressed on the sum of the probabilities. That is, for the k-th crop, $Y_{ki.}$ is regressed on $Z_{ki.}$ with an intercept, where

$Y_{ki.}$ = the ground acreage of crop k in segment i,

$$Z_{ki.} = \sum_{j} \hat{p}(\theta=k|\underset{\sim}{X}_{ij}), \tag{4.A.2}$$

and the summation is over all pixels in segment i.

B. Criteria for Comparisons

Two regressions were computed to evaluate the alternative auxiliary variables. The first regression used the individual pixels as observations. In this case, the dependent variable for crop j is one if the ground truth of the pixel is crop j and zero otherwise. The independent variables are the values of the auxiliary variables for the

pixel. In the second regression, the dependent variable for crop j is the acres of crop j in the segment and the independent variables are the sum of the values of the auxiliary variables for all the pixels in the segment. As demonstrated in Chapter III, the effect due to the estimation in auxiliary variables can be neglected for large samples provided that the auxiliary variables satisfy certain conditions. However, discrete variables, such as the USDA pixel classifier, may not satisfy those imposed conditions. To make comparisons, we assume that the effect due to the estimation in the discrete auxiliary variables is also negligible for large samples. Under such assumptions, either the R^2 values or the residual mean square errors were used for making comparisons among various estimated auxiliary variables.

C. Normal Class-conditional Probability

We assume that the vector $\underset{\sim}{X}$, conditional on $\theta = k$ $(k=1,2,\ldots, 8)$ is distributed as a multivariate normal. With this assumption, the normal conditional distribution can be estimated by estimating the mean and covariance matrix for each crop. That is, for each k, the probability density of $\underset{\sim}{X}$ is estimated by

$$\hat{p}(\underset{\sim}{X}|\,\theta=k) = (2\pi)^{-2}|\hat{\underset{\sim}{\Sigma}}_k|^{-1/2}\exp[-0.5(\underset{\sim}{X} - \bar{\underset{\sim}{X}}_k)'\hat{\underset{\sim}{\Sigma}}_k^{-1}(\underset{\sim}{X} - \bar{\underset{\sim}{X}}_k)] \,, \tag{4.C.1}$$

where $\bar{\underset{\sim}{X}}_k$ is the sample mean vector of $\underset{\sim}{X}$ for crop k, $\hat{\underset{\sim}{\Sigma}}_k$ is the sample covariance matrix of $\underset{\sim}{X}$ for crop k; i.e.,

$$\bar{\underset{\sim}{X}}_k = \left[\sum_i \sum_j Y_{kij}\right]^{-1} \sum_i \sum_j Y_{kij}\underset{\sim}{X}_{ij} , \qquad (4.C.2)$$

$$\hat{\underset{\sim}{\Sigma}}_k = \left[\sum_i \sum_j Y_{kij} - 1\right]^{-1} \sum_i \sum_j Y_{kij}(\underset{\sim}{X}_{ij} - \bar{\underset{\sim}{X}}_k)'(\underset{\sim}{X}_{ij} - \bar{\underset{\sim}{X}}_k) . \qquad (4.C.3)$$

Note that in these calculations the fact that the data are clustered into segments is ignored. The estimators are ratio estimators and this fact must be recognized if variances of the estimators are calculated.

Let $\hat{f}_k$ be the sample fraction of the area in crop k. Then for each k, the posterior probability is estimated by $\hat{p}(\theta=k|\underset{\sim}{X})$ defined in (4.A.1).

The sample mean and covariance matrix were first computed for each crop using the sample of 19,943 pixels. Plotting of the data was sufficient to establish that the marginal distributions are not normal. To investigate the effect of alternative estimates of the parameters, all pixels with at least one coordinate of $\underset{\sim}{X}$ falling outside 2.5 standard deviations from the mean were screened. The mean, variances and conditional probabilities were then computed for the resulting sample of 18,907 pixels.

The R^2 values for pixel regressions of Y_{kij} on various independent variables are reported in Table 4.C.1. The regressions in the first column of Table 4.C.1 are the regression of Y_{kij} on $\hat{p}(\theta=k|\underset{\sim}{X})$ for the original 19,943 pixels with an intercept included in the regression. The R^2 values reported in the second column are

Table 4.C.1 - R-square values for pixel regressions on probabilities

	Original sample $\hat{p}$		Screened sample $\hat{p}$	
Crop	Own probability	Multiple	Own probability	Multiple
Corn	0.19	0.20	0.20	0.21
Soybeans	0.43	0.43	0.43	0.43
Pasture	0.17	0.17	0.17	0.17
Woods	0.12	0.13	0.13	0.13
Winter Wheat	0.09	0.09	0.08	0.08
Alfalfa	0.01	0.01	0.01	0.01
Sorghum	0.01	0.01	0.01	0.01
All other	0.05	0.05	0.04	0.05

obtained from the multiple regression of Y_{kij} on all eight $\hat{p}(\theta=k|\underset{\sim}{X})$ with no intercept in the regression. (The sum of the eight $\hat{p}(\theta=k|\underset{\sim}{X})$ is one.) One can easily see that there is no significant improvement in the R^2 associated with the addition of the other conditional probabilities to each crop model. Note that these R^2 values are calculated on a pixel basis. These results tend to support the normal model because the other probabilities might improve the R^2 values for nonnormal distributions. The R^2 values in the two columns under "Screen" are for the same two regressions computed for the sample of 19,943 pixels, but using the $\hat{p}(\theta=k|\underset{\sim}{X})$ computed with the estimates obtained from the screened sample of 18,907 pixels. The use of the screened sample to construct the conditional probabilities had little

effect upon the correlations in the pixel regressions. The correlations for corn and woods were slightly higher for probabilities constructed from screened data while the correlations for winter wheat and all other were slightly smaller. On the basis of these results, one tentatively concludes that screening as an adjustment for nonnormality is not worthwhile.

Let a G-variable and an H-variable be defined by

$$G_{kij} = \begin{cases} 1 \text{ if } \hat{p}(\theta=k|\underset{\sim}{X}_{ij}) = \max(S_{ij}) \\ 0 \text{ otherwise,} \end{cases}$$

$$H_{kij} = \begin{cases} 1 \text{ if } \hat{p}(\theta=k|\underset{\sim}{X}_{ij}) = \max(S^*_{ij}) \\ 0 \text{ otherwise,} \end{cases}$$

where

$$S_{ij} = \{\hat{p}(\theta=1|\underset{\sim}{X}_{ij}),\ldots, \hat{p}(\theta=8|\underset{\sim}{X}_{ij})\}\ ,\quad j = 1,2,\ldots, M_i,\ i = 1,2,\ldots, N\ ,$$

S^*_{ij} = set S_{ij} with maximum probability deleted .

Let

$G_{ki.} = \sum_j G_{kij}$ = number of pixels classified as crop k in segment i by the rule based on G .

Thus, G_{kij} is one for the ij-th pixel if the k-th conditional probability is the largest for that pixel and H_{kij} is one for the ij-th pixel if the k-th conditional probability is the second largest for that pixel. The variable G provides a method of assigning each pixel to a crop group. Table 4.C.2 provides the number of pixels classified for each crop by this maximum probability classification rule G . One can see that there is a sharp contrast between the sample frequency shown in the first column and the frequency induced by the maximum probability classification rule in the second column. The G-rule classifies more pixels as pasture than were observed, whereas no pixels are classified into alfalfa or sorghum. This comparison shows that the G-rule is not suitable for rare crops.

Table 4.C.2 - Sample frequencies and frequencies induced by G-rule

Crop	Frequency	Number of pixels classified by maximum probability rule G
Corn	1,836	2,152
Soybeans	4,547	5,069
Pasture	6,710	11,099
Woods	1,567	290
Winter Wheat	460	139
Alfalfa	217	0
Sorghum	145	0
All other	4,431	1,194

Table 4.C.3 contains the R^2 values for pixel regressions containing G, H, and P as independent variables, where $P = \hat{p}(\theta=k|\underset{\sim}{X}_{ij})$ is computed using all 19,943 pixels. The R^2 for the regression on P is uniformly higher than that based on G. It is also uniformly higher than the regression using the USDA classification. The regression R^2 for G is comparable to that for the USDA classification. In no case did the addition of G and H to the regression containing P produce a significant improvement at the 5 percent level. The USDA classification is nonlinear and is based upon pixels interior to a field. Also, the R^2-value for the USDA procedure is maximized by varying the prior probabilities. Therefore, one concludes that the R^2 for the rule based on G would be slightly superior to that of the USDA classification if both used the same probability distributions.

The regressions in the last column of Table 4.C.3 are the multiple regression of Y_{kij} on all the linear and quadratic effects of $\underset{\sim}{X}_{ij}$ (i.e., X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_3^2, X_4^2, X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2,X_4, and X_3X_4) with an intercept. The R^2 values reveal that the linear model with all X_i's and all X_iX_j's as independent variables performs nearly as well as the model with conditional probabilities for pixel regressions.

The segment regressions computed using probability sums as independent variables and the segment crop acreages as dependent variables are compared with regressions using acreages obtained by the

Table 4.C.3 - R-square values for pixel regressions - original data

Crop	Independent variables in the regressions						
	USDA classification	G	G,H	G,P	P	G,H,P	all X_i, all X_iX_j
Corn	0.12	0.14	0.17	0.19	0.19	0.19	0.15
Soybeans	0.36	0.37	0.38	0.43	0.43	0.43	0.42
Pasture	0.12	0.12	0.14	0.17	0.17	0.17	0.14
Woods	0.08	0.05	0.08	0.12	0.12	0.13	0.10
Winter Wheat	*	0.05	0.06	0.09	0.09	0.09	0.06
Alfalfa	*	0.00	0.00	0.01	0.01	0.01	0.01
Sorghum	*	0.00	0.00	0.01	0.01	0.01	0.01
All other	*	0.02	0.02	0.05	0.05	0.05	0.05

USDA pixel classification scheme as independent variables in Table 4.C.4. On the basis of the R-square values, the posterior probabilities are no better than the USDA classification acres as auxiliary variables for segment data. This was somewhat surprising because the conditional probabilities performed much better in the pixel regressions. The heavy clustering of the data seems to negate the improvement obtained at the pixel level. Also, the maximizing of the correlation with respect to the prior probabilities used by USDA improves the apparent performance of the USDA classifier. As in the pixel regressions, the probability sums, based upon screened data, sometimes performed slightly

Table 4.C.4 - R-square values for segment regressions:

Crop	USDA classi-fication	Multiple regr. prob.	Prob. sum	Prob. sum (screened)	X-quadratic prediction
Corn	0.13	0.05	0.04	0.07	0.01
Soybeans	0.85	0.80	0.81	0.83	0.76
Pasture	0.86	0.86	0.86	0.87	0.84
Woods	0.46	0.33	0.29	0.35	0.24
Winter Wheat	*	0.08	0.08	0.06	0.01
Alfalfa	*	0.02	0.02	0.02	0.01
Sorghum	*	0.13	0.14	0.17	0.04
All other	*	0.34	0.33	0.29	0.33

better than and sometimes slightly worse than the probabilities estimated from the original data.

The R^2 values in the second column of Table 4.C.4 are obtained for the segment regressions with independent variable equal to the sum of the predicted values (Y-hats) calculated from the pixel multiple regression of Y_{kij} on all eight $\hat{p}(\theta=k|\underset{\sim}{X})$. Based on the R^2 values, the sum of multiple regression predicted values performs no better than the own probability sum. This is not surprising because there is no significant improvement in the R^2 value of the pixel regressions as other conditional probabilities are added to the pixel regression model for each crop. Finally, to investigate the performance of using all X_i's and all X_iX_j's as auxiliary variables, the predicted values were

obtained from the pixel regression of Y_{kij} on all X_i's and all X_iX_j's with an intercept. The sums of the predicted values for the segment served as the independent variable in the segment regressions whose R^2 values are reported in the last column of Table 4.C.4. These X-quadratic-prediction values generally give lower R^2 values than the probability sums. The differences are roughly comparable to the differences in the pixel R^2 values reported in Table 4.C.3.

In this particular scene, the $\underset{\sim}{X}$-distribution is very similar for corn and woods. If we add the probability sum for woods to the corn regression, the R^2 increases to 0.57. The R^2 for the analogous multiple regression using the two USDA classifications is 0.19. If the probability sum for corn is added to the woods regression, the R^2 is 0.29. The R^2 for the multiple regression of woods on USDA classified woods and corn is 0.46. Corn is the only cover type for which a very significant improvement was obtained by adding a second probability sum or USDA classification to the respective regressions.

D. The Distribution Function with Normal Conditional Probability

In this section, we develop an alternative distribution function for use in estimating the conditional probabilities. Plotting of the data revealed two types of nonnormality. First, the data were often skewed with a few observations large (or small) relative to the mean. Data transformations and screening of the type described in the preceding sections are methods of treating this problem. Second, it was noted that the conditional mean of X_i given (X_j, X_k, X_ℓ) was not

always a linear function of (X_j, X_k, X_ℓ) . Methods of recognizing the second type of nonnormality are considered in this section. For each $i = 1,2,\ldots, k$, let θ_i denote the event $\theta = i$. We assume

(1) $p(X_1|\theta_i) \sim N(\mu_i, \sigma_i^2)$.

(2) $p(X_2|X_1; \theta_i)$ is a normal density with mean $E(X_2|X_1; \theta_i) = f_i(X_1)$ and variance $V(X_2|X_1; \theta_i) = \xi_i^2$.

(3) $p(X_3|X_2, X_1; \theta_i)$ is a normal density with mean $E(X_3|X_2, X_1; \theta_i) = g_i(X_1, X_2)$, and variance $V(X_3|X_1, X_2; \theta_i) = \gamma_i^2$.

(4) $p(X_4|X_3, X_2, X_1; \theta_i)$ is a normal density with mean $E(X_4|X_1, X_2, X_3; \theta_i) = h_i(X_1, X_2, X_3)$, and variance $V(X_4|X_1, X_2, X_3; \theta_i) = \lambda_i^2$.

ξ_i^2, γ_i^2 , and λ_i^2 are assumed to be constant in X_1, X_2 , and X_3 . It would be possible to modify the model to permit ξ_i^2, γ_i^2 and λ_i^2 to be functions of X_1, (X_1, X_2) and (X_1, X_2, X_3) , respectively. Then, the joint density of $\underset{\sim}{X}$ for crop i is

$$p(\underset{\sim}{X}|\theta_i) = p(X_1|\theta_i)p(X_2|X_1; \theta_i)p(X_3|X_2X_1; \theta_i)p(X_4|X_1, X_2, X_3; \theta_i)$$

$$= (2\pi)^{-2}(\sigma_i\xi_i\gamma_i\lambda_i)^{-1}\exp(-\tfrac{1}{2}\{(\sigma_i^{-2}(X_1 - \mu_i)^2 + \xi_i^{-2}[X_2 - f_i(X_1)]^2$$

$$+ \gamma_i^{-2}[X_3 - g_i(X_1, X_2)]^2 + \lambda_i^{-2}[X_4 - h_i(X_1, X_2, X_3)]^2\}) ,$$

$$i = 1,2,\ldots, 8 . \qquad (4.D.1)$$

Note that if $p(\underset{\sim}{X}|\theta_i)$ is multivariate normal, then all of the above assumptions hold, and all of the above mean functions are linear.

Regression analysis of the sample of 45 segments demonstrated that the mean functions are not linear in X_1, X_2, and X_3. For example, the pixel corn regression gave

$$\hat{X}_2 = \underset{(3.6)}{50.4} - \underset{(0.41)}{4.91}\, X_1 + \underset{(0.011)}{0.164}\, X_1^2 .$$

The t-value for the X_1^2 effect is 14.28 and this t-value suggests that the quadratic effect should be included in the mean function. On the other hand, the residual plots indicated that the assumption of constant variance is acceptable.

We started the analysis with the screened data for X_1. The resulting sample is then successively screened when the absolute values of the standardized residuals from the regressions are greater than 2.5. For instance, at the second stage, observations are screened when the absolute values of the standardized residuals from the regression of X_2 on X_1, X_1^2 with intercept exceed 2.5.

The mean functions for each crop were estimated for the sample of 18,562 pixels. The quadratic variable was included in the mean function when the t-value was greater than 2.0 in absolute value. The residual mean square error is an estimate of the conditional variance. Then the

class-conditional probability of $\underset{\sim}{X}$ for crop i was estimated by substituting the estimated parameters into (4.D.1). For example, for corn,

$$\hat{f}_1(X_1) = \hat{X}_2 = \underset{(3.2)}{51.8} - \underset{(0.36)}{5.05}\, X_1 + \underset{(0.010)}{0.167}\, X_1^2 \quad , \qquad \hat{\xi}_1^2 = 1.14830$$

$$\hat{g}_1(X_1, X_2) = \hat{X}_3 = \underset{(16.)}{84.} - \underset{(1.7)}{3.9}\, X_1 - \underset{(0.11)}{0.32}\, X_2 + \underset{(0.050)}{0.138}\, X_1^2 \quad , \quad \hat{\gamma}_1^2 = 24.1318$$

$$\hat{h}_1(X_1,X_2,X_3) = \hat{X}_4 = \underset{(9.2)}{12.5} - \underset{(0.68)}{2.77}\, X_1 - \underset{(0.72)}{0.31}\, X_2 + \underset{(0.20)}{1.82}\, X_3 + \underset{(0.015)}{0.062}\, X_2^2$$

$$- \underset{(0.0018)}{0.0084}\, X_3^2 + \underset{(0.013)}{0.047}\, X_1X_3 - \underset{(0.010)}{0.042}\, X_2X_3$$

$$\hat{\lambda}_1^2 = 6.31683 \ ,$$

$$\hat{\sigma}_1^2 = 1.72815 \ ,$$

$$\bar{X}_1 = 17.4 \ .$$

The posterior probability computed in this way is superior to the USDA classification variable in terms of the R^2 values for the pixel regressions. However, as shown in Table 4.D.1, the R^2 values are not much different from those obtained using the conditional probabilities based upon standard normal density. Table 4.D.2 contains the R^2 for the segment regressions. The probability sum for the conditionally normal distribution performs marginally better than the sum based on the unconditional normal approximation. While not important in this example, this improvement illustrates that the method of estimating the posterior probability may play an important role.

Table 4.D.1 - R^2 values for pixel regression

Crop	Normal	Nonlinear mean	USDA classifier
Corn	0.20	0.20	0.12
Soybeans	0.43	0.43	0.36
Pasture	0.17	0.18	0.12
Woods	0.13	0.13	0.08
Winter Wheat	0.08	0.10	*
Alfalfa	0.01	0.00	*
Sorghum	0.01	0.01	*
All other	0.04	0.05	*

Table 4.D.2 - R^2 values for segment regression

Crop	USDA	Normal conditional	Normal Screened	Normal original
Corn	0.13	0.08	0.07	0.04
Soybeans	0.85	0.83	0.83	0.81
Pasture	0.86	0.88	0.87	0.86
Woods	0.46	0.37	0.35	0.29
Winter Wheat	*	0.10	0.06	0.08
Alfalfa	*	0.06	0.02	0.02
Sorghum	*	0.17	0.17	0.14
All other	*	0.33	0.29	0.33

E. Restricted Segment Multiple Regression

In this section, we apply the generalized least squares to the estimation of the system of segment regressions for all crops. Let

$$D_t = \sum_{j=1}^{8} p(\underset{\sim}{X}_t | \theta=j) f_j$$

and

$$g_{ti} = D_t^{-1} \, p(\underset{\sim}{X}_t | \theta=i) \ .$$

If f_j is the proportion in the population, then

$$E\{n^{-1} \sum_{t=1}^{n} g_{ti}\} = 1 \ , \quad i = 1,2,\ldots, 8$$

for a random sample of size n .

For a random sample, if $p(\underset{\sim}{X}_t | \theta=j)$, $j = 1,2,\ldots, 8$, are known, the maximum likelihood estimator of f_j , denoted by $\hat{f}_j$, will satisfy

$$n^{-1} \sum_{t=1}^{n} \hat{g}_{ti} = 1 \ , \quad i = 1,2,\ldots, 8 \ , \qquad (4.E.1)$$

where

$$\hat{D}_t = \sum_{j=1}^{8} p(\underset{\sim}{X}_t | \theta=j) \hat{f}_j \ ,$$

$$\hat{g}_{ti} = \hat{D}_t^{-1} \; p(\underset{\sim}{X}_t \,|\, \theta=i) \; .$$

Assume that we have initial estimates of f_j, denoted by $\tilde{f}_j$, and that we wish to obtain a set of f_j satisfying (4.E.1). Because $\sum_{j=1}^{8} \hat{f}_j = \sum_{j=1}^{8} \tilde{f}_j = 1$, we can impose the condition

$$\sum_{j=1}^{8} (\hat{f}_j - \tilde{f}_j) = 0 \; .$$

By using the restriction that the sum of the changes must be zero, one way to proceed is to expand (4.E.1) in a Taylor series about $\tilde{f}_i$ to obtain

$$n^{-1} \sum_{t=1}^{n} (\tilde{g}_{ti} - \tilde{g}_{t8}) - n^{-1} \sum_{t=1}^{n} (\tilde{g}_{ti} - \tilde{g}_{t8}) \sum_{j=1}^{7} (\tilde{g}_{tj} - \tilde{g}_{t8})(\hat{f}_j - \tilde{f}_j) = 0 \; , \quad i = 1,2,\ldots, 7 \; . \qquad (4.E.2)$$

The system (4.E.2) can be written as

$$\sum_{j=1}^{7} C_{ij}(\Delta f_j) = v_i \; , \quad i = 1,2,\ldots, 7 \; ,$$

where

$$C_{ji} = C_{ij} = n^{-1} \sum_{t=1}^{n} (\tilde{g}_{ti} - \tilde{g}_{t8})(\tilde{g}_{tj} - \tilde{g}_{t8}) \; ,$$

$$v_i = n^{-1} \sum_{t=1}^{n} (\tilde{g}_{ti} - \tilde{g}_{t8}) ,$$

$$\Delta f_j = \hat{f}_j - \tilde{f}_j .$$

Solving the system for the change in the estimated fractions, we have

$$\underset{\sim}{\Delta f} = \underset{\sim}{B}\, \underset{\sim}{v} , \qquad (4.E.3)$$

where

$$\underset{\sim}{B} = \underset{\sim}{C}^{-1} ,$$

$$\underset{\sim}{\Delta f} = (\Delta f_1, \Delta f_2, \ldots, \Delta f_7)' ,$$

$$\underset{\sim}{v} = (v_1, v_2, \ldots, v_7)'$$

and $\underset{\sim}{C}$ is the matrix with elements C_{ij}. We shall construct an estimator of $\underset{\sim}{B}$. An estimator of the difference between the fraction of crop j in the ℓ-th segment and the population fraction is given by

$$d_{\ell j} = m_\ell^{-1} a_{\ell j} - \tilde{f}_j ,$$

where m_ℓ is the number of pixels in segment ℓ and $a_{\ell j}$ is the number of pixels identified as crop j on the basis of ground truth. The $d_{\ell j}$ are sample segment equivalents of Δf_j. The quantity

$$v_{\ell j} = m_\ell^{-1} \sum_{t=1}^{m_\ell} (\tilde{g}_{\ell t j} - \tilde{g}_{\ell t 8}) ,$$

where

$$\tilde{g}_{\ell t j} = [\sum_{i=1}^{8} \hat{p}(\underset{\sim}{X}_{\ell t} | \theta=i)\tilde{f}_i]^{-1} \hat{p}(\underset{\sim}{X}_{\ell t} | \theta=i) , \qquad (4.E.4)$$

is a sample segment equivalent to v_j. Then, viewing (4.E.3) as a regression equation we can estimate the coefficients B_{ij} by the regression of $d_{\ell j}$ on $(v_{\ell 1}, v_{\ell 2}, \ldots, v_{\ell 7})$ or (using sums) by the regression of $m_\ell d_{\ell j}$ on $(m_\ell v_{\ell 1}, m_\ell v_{\ell 2}, \ldots, m_\ell v_{\ell 7})$. Because the matrix $\underset{\sim}{B}$ is symmetric, the regression equations should be estimated subject to this restriction. Also, the errors in the regression equations will be correlated across equations. To estimate these equations, we used the "3-stage" or "Seemingly Unrelated" option of SAS. This is a type of generalized least squares that recognizes the correlation structure in the estimation. Given the regression coefficients, the estimator of the mean acres of crop j per segment is

$$\bar{Y}_{j(reg)} = \bar{y}_j + \sum_{i=1}^{7} \hat{B}_{ji}(\bar{T}_{i(p)} - \bar{T}_{i(s)}) , \quad j = 1,2,\ldots, 7 ,$$

$$\bar{Y}_{8(reg)} = \bar{y}_8 + \sum_{i=1}^{7} (- \sum_{j=1}^{7} \hat{B}_{ji})(\bar{T}_{i(p)} - \bar{T}_{i(s)}) , \qquad (4.E.5)$$

where

$\bar{T}_{\ell i(p)}$ is the population mean per segment of v_i ,

$\bar{T}_{\ell i(s)}$ is the sample mean per segment of v_i ,

and $\hat{B}_{ji}$ are the regression estimates of B_{ji} . The estimators are multiple regression estimators with the coefficients restricted by certain linear constraints. On the other hand, viewing (4.E.3) as a possible model for $\Delta\underset{\sim}{f}$, we can then develop a more appropriate model for $\Delta\underset{\sim}{f}$. Note that the vector of ones is not in the column space spanned by the design matrix in each regression equation of system (4.E.3). Therefore, we expect that intercept may need to be included in the regressions. On the other hand, the correlational analysis revealed that Δf_3 seems to be correlated to segment acres. The linear correlation between Δf_3 and the reciprocal of segment acres, M^{-1} , is slightly higher than that between Δf_3 and segment acres. Hence, with an intercept and M^{-1} included in all the regressions, the full model for $\Delta\underset{\sim}{f}$ becomes

$$\hat{d}_1 = \underset{(0.072)}{-\ 0.071} + \underset{(0.030)}{0.188}\ v_1 \underset{(0.011)}{-\ 0.013}\ v_2 \underset{(0.030)}{-\ 0.116}\ v_3$$

$$- \underset{(0.014)}{0.005}\, v_4 - \underset{(0.030)}{0.095}\, v_5 - \underset{(0.0097)}{0.0044}\, v_6 - \underset{(0.012)}{0.012}\, v_7$$

$$+ \underset{(30.)}{30.}\, M^{-1} ,$$

$$\hat{d}_2 = \underset{(0.035)}{-\,0.017} - \underset{(0.011)}{0.013}\, v_1 + \underset{(0.0086)}{0.0233}\, v_2 - \underset{(0.017)}{0.020}\, v_3$$

$$+ \underset{(0.0076)}{0.0054}\, v_4 + \underset{(0.013)}{0.018}\, v_5 + \underset{(0.0069)}{0.0019}\, v_6 + \underset{(0.0087)}{0.0026}\, v_7$$

$$+ \underset{(14.)}{4.}\, M^{-1}$$

$$\hat{d}_3 = \underset{(0.12)}{0.31} - \underset{(0.030)}{0.116}\, v_1 - \underset{(0.017)}{0.020}\, v_2 + \underset{(0.065)}{0.417}\, v_3 - \underset{(0.024)}{0.116}\, v_4$$

$$- \underset{(0.035)}{0.050}\, v_5 - \underset{(0.019)}{0.026}\, v_6 + \underset{(0.025)}{0.016}\, v_7 - \underset{(51.)}{129.}\, M^{-1} ,$$

$$\hat{d}_4 = \underset{(0.086)}{-\,0.056} - \underset{(0.014)}{0.005}\, v_1 + \underset{(0.0076)}{0.0054}\, v_2 - \underset{(0.024)}{0.116}\, v_3$$

$$+ \underset{(0.016)}{0.195}\, v_4 - \underset{(0.016)}{0.038}\, v_5 - \underset{(0.0081)}{0.0030}\, v_6 - \underset{(0.012)}{0.029}\, v_7$$

$$+ \underset{(36.)}{18.}\, M^{-1} ,$$

$$\hat{d}_5 = \underset{(0.084)}{-\,0.059} - \underset{(0.030)}{0.095}\, v_1 + \underset{(0.013)}{0.018}\, v_2 - \underset{(0.035)}{0.050}\, v_3$$

$$- \underset{(0.016)}{0.038}\, v_4 + \underset{(0.039)}{0.176}\, v_5 + \underset{(0.012)}{0.016}\, v_6 + \underset{(0.015)}{0.014}\, v_7$$

$$+ \underset{(35.)}{27.}\ M^{-1} ,$$

$$\hat{d}_6 = \underset{(0.025)}{0.023} - \underset{(0.0097)}{0.0044}\ v_1 + \underset{(0.0069)}{0.0019}\ v_2 - \underset{(0.019)}{0.026}\ v_3$$

$$- \underset{(0.0081)}{0.0030}\ v_4 + \underset{(0.012)}{0.016}\ v_5 + \underset{(0.022)}{0.061}\ v_6 - \underset{(0.016)}{0.006}\ v_7$$

$$- \underset{(10.)}{10.}\ M^{-1} ,$$

$$\hat{d}_7 = \underset{(0.033)}{0.041} - \underset{(0.015)}{0.012}\ v_1 + \underset{(0.016)}{0.0026}\ v_2 + \underset{(0.023)}{0.016}\ v_3$$

$$- \underset{(13.)}{15.}\ M^{-1} .$$

If the variables with t-values greater than 2.0 are included in the regression, one can see that only for pasture are the intercept and the reciprocal of the segment acres included in the model. The reduced model for $\Delta \underset{\sim}{f}$ becomes

$$\hat{d}_1 = \underset{(0.026)}{0.179}\ v_1 - \underset{(0.019)}{0.134}\ v_3 - \underset{(0.027)}{0.080}\ v_5 ,$$

$$\hat{d}_2 = \underset{(0.0068)}{0.0222}\ v_2 ,$$

$$\hat{d}_3 = \underset{(0.093)}{0.287} - \underset{(0.019)}{0.134}\ v_1 + \underset{(0.055)}{0.388}\ v_3 - \underset{(0.019)}{0.110}\ v_4 - \underset{(39.)}{124.}\ M^{-1} ,$$

$$\hat{d}_4 = \underset{(0.019)}{-0.110}\ v_3 + \underset{(0.015)}{0.188}\ v_4 - \underset{(0.010)}{0.036}\ v_5 - \underset{(0.010)}{0.023}\ v_7 ,$$

$$\hat{d}_5 = \underset{(0.027)}{-0.080}\, v_1 - \underset{(0.010)}{0.036}\, v_4 + \underset{(0.034)}{0.146}\, v_5 ,$$

$$\hat{d}_6 = \underset{(0.014)}{0.036}\, v_6 ,$$

$$\hat{d}_7 = \underset{(0.010)}{-0.023}\, v_4 + \underset{(0.016)}{0.059}\, v_7 .$$

If we multiply $\Delta\underset{\sim}{f}$ and $\underset{\sim}{y}$ by the segment acres in system (4.E.3), the resulting system of regression equations can be viewed as a possible model for $\Delta\underset{\sim}{f}$ multiplied by segment acres. In the segment analysis for corn, winter wheat, and pasture, intercept and segment acres become significant in the model.

Based on the same selection procedure, the reduced model for $\Delta\underset{\sim}{f}$ multiplied by the segment acres is

$$\hat{e}_1 = \underset{(12.)}{68.} + \underset{(0.027)}{0.187}\, t_1 - \underset{(0.019)}{0.125}\, t_3 - \underset{(0.027)}{0.096}\, t_5 - \underset{(0.027)}{0.160}\, M ,$$

$$\hat{e}_r = \underset{(7.7)}{19.4} + \underset{(0.0066)}{0.0238}\, t_2 - \underset{(0.017)}{0.050}\, M ,$$

$$\hat{e}_3 = \underset{(23.)}{-102.} - \underset{(0.019)}{0.125}\, t_1 + \underset{(0.050)}{0.385}\, t_3 - \underset{(0.018)}{0.116}\, t_4 + \underset{(0.050)}{0.234}\, M ,$$

$$\hat{e}_4 = \underset{(0.017)}{-0.116}\, t_3 + \underset{(0.014)}{0.195}\, t_4 - \underset{(0.0097)}{0.0282}\, t_5 - \underset{(0.0099)}{0.0197}\, t_7 ,$$

$$\hat{e}_5 = \underset{(0.027)}{-0.096}\, t_1 - \underset{(0.0097)}{0.0282}\, t_4 + \underset{(0.034)}{0.165}\, t_5 ,$$

$$\hat{e}_6 = \underset{(0.012)}{0.038}\ t_6 \ ,$$

$$\hat{e}_7 = \underset{(0.0099)}{-0.0197}\ t_4 + \underset{(0.015)}{0.049}\ t_7 \ ,$$

where M is segment acres and

$$e_i = d_i M \ ,$$

$$t_i = v_i M \ , \quad i = 1,2,\ldots, 7 \ .$$

F. Comparisons

In previous sections, we considered constructing the regression estimators of mean acreage per segment (or total acres) for each crop by

Table 4.F.1 - Sample linear correlation

Crop	Ground-truth		Probability sum	
Corn	-0.16	(0.30)	0.63	(0.0001)
Soybeans	-0.11	(0.45)	-0.01	(0.94)
Pasture	0.75	(0.0001)	0.78	(0.0001)
Woods	0.17	(0.26)	0.61	(0.0001)
Winter Wheat	-0.08	(0.60)	0.63	(0.0001)
Alfalfa	-0.03	(0.87)	0.61	(0.0001)
Sorghum	-0.02	(0.92)	0.04	(0.78)
All other	0.59	(0.0001)	0.85	(0.0001)

using the posterior probabilities. In Table 4.F.1, we report the sample linear correlations between ground-truth acres and segment acres in the first column, and the correlations between probability sums and segment acres in the second column. The numbers in the parentheses are the probabilities of a test statistic as large as or larger than the absolute values of the corresponding correlations under the hypothesis that the population linear correlation is zero.

From these correlational analyses, the segment acres may have a significant effect on the regression used for the estimation of crop acres. One simple method of taking the segment acres into account in the regression estimation is to include the segment acres as an independent variable in the regression equations. In this section, we make comparisons among several possible estimators of which some are constructed with the segment acres added to the segment regressions.

We repeat the definitions of the variables:

M_k = number of pixels in segment k ,

$$G_{kij} = \begin{cases} 1 & \text{if } \hat{p}(\theta=k|\underset{\sim}{X}_{ij}) = \max\{\hat{p}(\theta=1|\underset{\sim}{X}_{ij}),\ldots, \hat{p}(\theta=8|\underset{\sim}{X}_{ij})\} \\ 0 & \text{otherwise,} \end{cases}$$

$G_{ki.}$ = number of pixels classified as crop k in segment i by the rule based on G ,

$W_{ki.}$ = number of pixels classified as crop k in segment i by the USDA pixel classification,

$Z_{ki.}$ = the sum of the estimated posterior probabilities for crop k in segment i,

$$t_{ki} = \Sigma_{j=1}^{M_k} (\tilde{g}_{kji} - \tilde{g}_{kj8}) ,$$

$$v_{ki} = M_k^{-1} t_{ki} ,$$

where $\tilde{g}_{kji}$ is defined in (4.E.4). Note that the maximum probabilty rule G can be viewed alternatively in the following way:

Define

$$R_{kij} = \frac{\hat{p}(\theta=k \mid \underset{\sim}{X}_{ij})}{\max\{\hat{p}(\theta=1 \mid \underset{\sim}{X}_{ij}), \ldots, \hat{p}(\theta=8 \mid \underset{\sim}{X}_{ij})\}} , \quad \text{for each } i,j,k . \tag{4.F.1}$$

Then,

$$G_{kij} = \begin{cases} 1 & \text{if } R_{kij} = 1 \\ 0 & \text{otherwise} \end{cases} .$$

As we mentioned in the previous sections, one deficiency for this G is that it tends to classify no pixels to the rare crops such as alfalfa and sorghum. To make up for this deficiency, a new G will be defined in a broader sense. One extension is to order the ratios R in (4.F.1) for all the 19,943 pixels and then set the value of R for the first 19,943 $\hat{f}$ pixels equal to one. The new rule G^* can be expressed as follows:

(i) If the crop k is such that $\sum_i \sum_j G_{kij} > f_k$,

$$G^*_{kij} = G_{kij} .$$

Table 4.F.2 - R-square values for pixel regressions

Crop	Nonlinear-mean probability	Maximum probability G^*
Corn	0.20	0.15
Soybeans	0.43	0.37
Pasture	0.17	0.13
Woods	0.13	0.09
Winter wheat	0.08	0.07
Alfalfa	0.01	0.001
Sorghum	0.01	0.004
All other	0.04	0.02

(ii) If the crop k is such that $\sum_i \sum_j G_{kij} < f_k$,

$$G^*_{kij} = \begin{cases} 1 & \text{if } R_{kij} \text{ of (4.F.1) is in the set of the 19,943} \\ & \hat{f}_k \text{ largest R-ratios} \\ 0 & \text{otherwise.} \end{cases}$$

Under this new "classification rule" G^* , the fraction of pixels "classified" into each crop is at least as large as the respective crop fraction. Denote $\sum_j G^*_{kij}$ by $G^*_{ki.}$ for each i,k .

Table 4.F.2 contains the R^2 values for pixel regressions containing the nonlinear-mean posterior probability and the maximum probability rule G^* as independent variables. The R^2 for the regression on the posterior probability is uniformly higher than the regression using the G^*-rule.

We denote by $\bar{Q}$ the average per segment for a variable Q . For notational convenience, let us suppress the subscript for crop type. Several possible estimators are:

$$\hat{\bar{Y}}_P = \bar{Y}_{(s)} + \hat{\beta}_{11}(\bar{Z}_{(p)} - \bar{Z}_{(s)}) ,$$

$$\hat{\bar{Y}}_{PM} = \bar{Y}_{(s)} + \hat{\beta}_{21}(\bar{Z}_{(p)} - \bar{Z}_{(s)}) + \hat{\beta}_{22}(\bar{M}_{(p)} - \bar{M}_{(s)}) ,$$

$$\hat{\bar{Y}}_{MRF} = \bar{Y}_{(s)} + \sum_{j \in I} \hat{\beta}_{3j}(\bar{t}_{j(p)} - \bar{t}_{j(s)}) + \hat{\beta}_{38}(\bar{M}_{(p)} - \bar{M}_{(s)}) ,$$

$$\hat{\bar{Y}}_{MRT} = \bar{Y}_{(s)} + \sum_{j\in I'} \hat{\beta}_{4j}(\bar{t}_{j(p)} - \bar{t}_{j(s)}) + \hat{\beta}_{48}(\bar{M}_{(p)} - \bar{M}_{(s)}) ,$$

$$\hat{\bar{Y}}_{W} = \bar{Y}_{(s)} + \hat{\beta}_{61}(\bar{W}_{(p)} - \bar{W}_{(s)}) ,$$

$$\hat{\bar{Y}}_{G*} = \bar{y}_{(s)} + \hat{\beta}_{71}(\bar{G}*_{(p)} - \bar{G}*_{(s)}) ,$$

$$\hat{\bar{Y}}_{PG*} = \bar{Y}_{(s)} + \hat{\beta}_{81}(\bar{Z}_{(p)} - \bar{Z}_{(s)}) + \hat{\beta}_{82}(\bar{G}*_{(p)} - \bar{G}*_{(s)}) ,$$

where

$\hat{\beta}_{rj}$ = the sample regression coefficient in the corresponding regression, $r = 1,2,3,\ldots, 8$, $j \neq 8$,

$$\hat{\beta}_{48} = \begin{cases} \text{sum of regression coefficient and sample fraction if } M \text{ is in the reduced model} \\ \text{sample fraction otherwise,} \end{cases}$$

$$\hat{\beta}_{38} = \begin{cases} \text{sum of intercept and sample fraction if intercept is significant} \\ \text{sample fraction otherwise,} \end{cases}$$

$I = \{j:\ j = 1,2,3,4,5,6,7;$ t-value of $\hat{\beta}_{3j}$ in the d model is greater than 2.0 in absolute value$\}$,

$I' = \{j:\ j = 1,2,,3,4,5,6,7;$ t-value of $\hat{\beta}_{4j}$ in the e model is is greater than 2.0 in absolute value$\}$,

and the suffixes (p) and (s) refer to the population and sample, respectively.

Table 4.F.3 contains the residual mean square error for each estimator. The degrees of freedom reported in the parentheses of the third and the fourth columns are obtained by the restricted segment multiple regression for the segment fraction model and segment total model, respectively. The coefficients estimated subject to the symmetry condition are assigned 0.5 degree of freedom rather than one. For instance, in the segment total model for corn the coefficients corresponding to corn, pasture, and woods have 0.5 degree of freedom each, while intercept and segment size have one degree of freedom each. For 'All other' as indicated in (4.E.5), the regression coefficients in the total acreages estimator are obtained from the restricted segment multiple regressions for the other seven crops. Therefore, the degrees of freedom for the residual mean squares tend to be between the two bounds of 36 and 39.5. The difference between these two bounds arises from the assignment of the degree of freedom to the sum of squares explained by the other seven crops in the restricted segment multiple regressions. The intercept and segment acres possess one degree of freedom each. The estimated variance for each estimator is the corresponding residual mean square error multiplied by a common fixed constant. One can see that the probability sum with segment acres, the probability sum with maximum probability classification rule G^* , and the restricted segment multiple regression procedure

based on the segment total, generally perform best. For corn, the residual mean square errors for the probability sum with maximum probability rule G* and the restricted segment multiple regression rule based on segment total model are about 60% of that for the probability sum with segment acres. It was surprising that for corn the residual mean square was reduced by about one half by the addition of the classification rule G* to the segment regression with probability sum as an independent variable. The improvement is about equivalent to the improvement obtained by the addition of the woods probability to the corn segment regression. However, unlike the situation where the addition of corn probability to the woods segment regression did not improve significantly the R^2 , the inclusion of the maximum probability rule G* in the woods regression reduces the woods residual mean square error by about 13%. Finally, the restricted segment multiple regresson procedure generally reduces the estimated variance for each crop except for 'all other'. When the variance of the crop total acreages estimate can be approximated by the linear term of a Taylor series expansion, the restricted segment multiple regression procedure based on the segment total model will be superior to that based on the segment fraction model in the sense that the former will give smaller estimated variance for the total crop acreages estimate. This is not surprising because the estimated coefficients in the segment total regression are the "best" in the sense of minimizing the residual sum of squares with respect to the segment total model.

Table 4.F.3 - Residual mean square error

Crop	Estimators P	PM	MRF	MRT	USDA	G*	PG*
Corn	2,073.	1,677.	1,506.(42)	1,035.(41)	1,976.	1,682.	1,013.
Soybeans	1,570.	1,507.	1,614.(41.5)	1,609.(41.5)	1,364.	1,518.	1,539.
Pasture	3,122.	3,178.	3,058.(41)	3,085.(41)	3,607.	3,161.	3,173.
Woods	1,366.	1,259.	1,485.(42)	1,456.(42)	1,165.	1,134.	1,160.
Winter Wheat	290.	259.	306.(43)	248.(42)	*	283.	285.
Alfalfa	123.	120.	117.(43)	117.(43)	*	124.	124.
Sorghum	207.	212.	204.(42.5)	200.(42.5)	*	209.	205.
All other	3,125.	3,042.	5,400.(39.5) 5,925.(36)	3,116.(39.5) 3,419.(36)	*	4,112.	3,110.

Numbers in the parentheses are corresponding degrees of freedom. Probability sums calculated using all data, but the class-conditional probabilities estimated based on 18,562 pixels.

P - Probability sum

PM - Probabililty sum and segment acres in multiple regression

MRF - Restricted segment multiple regression based on segment fraction model

MRT - Restricted segment multiple regression based on segment total model

USDA - USDA classification

G* - Maximum Probability Classification

PG* - Probability Sum and Maximum Probability Classification

G. Variance Estimation

In this section, we present various methods of estimating the variance of the regression estimator, when the regression estimator is constructed by using estimated probability sum as an auxiliary variable. To investigate how much the usual variance estimator underestimates the true variance, we compare a form of the Jackknife estimator with the usual variance estimator.

Recall that $\underset{\sim}{\mu}_k$ denotes the mean vector, $\underset{\sim}{\Sigma}_k$ denotes the positive definite covariance matrix, and f_k denotes the fraction of the area in crop k, $k = 1,2,\ldots, 8$. Let

$$Y_{ki} = \text{the ground acreage of crop } k \text{ in segment } i ,$$

$$Z_{ki} = \sum_{j=1}^{M_i} \hat{p}(\theta=k|\underset{\sim}{X}_{ij}) ,$$

where $\hat{f}_k$ is the sample fraction of the area in crop k, and the estimators, $\bar{\underset{\sim}{X}}_k$, $\hat{\underset{\sim}{\Sigma}}_k$, are defined in (4.B.1) - (4.B.2),

$$\hat{p}(\theta=k|\underset{\sim}{X}_{ij}) = \{\sum_{\ell=1}^{8} \hat{f}_\ell|\hat{\underset{\sim}{\Sigma}}_\ell|^{-1/2} \exp[-1/2\,(\underset{\sim}{X}_{ij} - \bar{\underset{\sim}{X}}_\ell)'\hat{\underset{\sim}{\Sigma}}_\ell^{-1}(\underset{\sim}{X}_{ij} - \bar{\underset{\sim}{X}}_\ell)]\}^{-1}$$

$$\times \hat{f}_k|\hat{\underset{\sim}{\Sigma}}_k|^{-1/2} \exp[-1/2\,(\underset{\sim}{X}_{ij} - \bar{\underset{\sim}{X}}_k)'\hat{\underset{\sim}{\Sigma}}_k^{-1}(\underset{\sim}{X}_{ij} - \bar{\underset{\sim}{X}}_k)] ,$$

$$M_i = \text{the number of pixels in segment } i .$$

Then, the sample average ground acreage per segment for crop k is $\bar{Y}_k^{(n)} = n^{-1} \Sigma_{i=1}^{n} Y_{ki}$, and the sample average estimated probability sum of crop k is $\bar{Z}_k^{(n)} = n^{-1} \Sigma_{i=1}^{n} Z_{ki}$.

To compute the Jackknife estimator, a set of 10 segments was randomly selected without replacement from the sample of 45 segments. One segment from the chosen set is deleted from the sample of 45 segments and the parameters are estimated using the remaining 44 segments. This operation is repeated 10 times. For notational convenience, let the chosen set contain the first 10 segments. Let

$Z_{ki(j)}$ = the sum of the estimated posterior probabilities for crop k in segment i, where all the parameters are estimated with segment j deleted,

$$\bar{Z}_{k(j)}^{(n)} = n^{-1} \sum_{i=1}^{n} Z_{ki(j)},$$

$$\bar{Z}_{k(j)}^{(n-1)} = (n-1)^{-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} Z_{ki(j)},$$

and

$$\bar{Y}_{k(j)}^{(n-1)} = (n-1)^{-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} Y_{ki}, \quad j = 1,2,\ldots, 10.$$

As segment j is deleted, a linear regression estimator of the mean population acreages per segment for crop k, $\bar{Y}_k^{(N)}$, can be obtained by estimating the population mean $\bar{Z}_k^{(N)}$ by $\bar{Z}_{k(j)}^{(n)}$ and is given by

$$\tilde{\bar{Y}}_{k(j)}^{(N)} = \bar{Y}_{k(j)}^{(n-1)} + \hat{\beta}_{k(j)}(\bar{Z}_{k(j)}^{(n)} - \bar{Z}_{k(j)}^{(n-1)}) ,$$

where

$$\hat{\beta}_{k(j)} = [\sum_{\substack{i=1 \\ i\neq j}}^{n} (Z_{ki(j)} - \bar{Z}_{k(j)}^{(n-1)})^2]^{-1} \sum_{\substack{i=1 \\ i\neq j}}^{n} (Z_{ki(j)} - \bar{Z}_{k(j)}^{(n-1)})(Y_{ki.} - \bar{Y}_{k(j)}^{(n-1)}) .$$

Note that $\hat{\beta}_{k(j)}$ is the sample regression coefficient obtained in a regression of Y_{ki} on $Z_{ki(j)}$ with an intercept using the 44 segments (segment j deleted). Define, for each j ,

$$\hat{\delta}_{kj}^2 = (\hat{\bar{Y}}_{k(j)}^{(N)} - \bar{Y}_{k}^{(n)})^2 ,$$

$$\delta_{kj}^2 = [Y_{kj} - \bar{Y}_{k}^{(n)} - \hat{\beta}_{k}(Z_{kj} - \bar{Z}_{k}^{(n)})]^2 ,$$

where

$$\hat{\beta}_{k} = [\sum_{i=1}^{n} (Z_{ki} - \bar{Z}_{k}^{(n)})^2]^{-1} \sum_{i=1}^{n} (Z_{ki} - \bar{Z}_{k}^{(n)})(Y_{ki} - \bar{Y}_{k}^{(n)}) .$$

One can see that $\hat{\delta}_{kj}^2 = (n-1)^{-2}[Y_{kj.} - \bar{Y}_{k}^{(n)} - \hat{\beta}_{k(j)}(Z_{kj(j)} - \bar{Z}_{k(j)}^{(n)})]^2$.
The usual variance estimator of the regression estimator $\hat{\bar{Y}}_{k}^{(N)}$, where

$$\hat{\bar{Y}}_k^{(N)} = \bar{Y}_k^{(n)} + \hat{\beta}_k(\bar{Z}_k^{(N)} - \bar{Z}_k^{(n)}) ,$$

is

$$v_{k0} = n^{-1} \hat{\sigma}_k^2 ,$$

where

$$\hat{\sigma}_k^2 = (n-2)^{-1} \sum_{i=1}^{n} \delta_{ki}^2 .$$

A ratio and a regression type estimator of the variance are presented. They are:

$$v_{k1} = c[10^{-1} \sum_{i=1}^{10} \hat{\delta}_{ki}^2 + \hat{\gamma}_k(45^{-1} \sum_{i=1}^{45} \delta_{ki}^2 - 10^{-1} \sum_{i=1}^{10} \delta_{ki}^2)] ,$$

$$v_{k2} = c[\sum_{i=1}^{10} \delta_{ki}^2]^{-1}[\sum_{i=1}^{10} \hat{\delta}_{ki}^2][45^{-1} \sum_{i=1}^{45} \delta_{ki}^2] ,$$

where

$$\hat{\gamma}_k = [\sum_{i=1}^{10} (\delta_{ki}^2 - 10^{-1} \sum_{i=1}^{10} \delta_{ki}^2)^2]^{-1} \sum_{i=1}^{10} (\delta_{ki}^2 - 10^{-1} \sum_{i=1}^{10} \delta_{ki}^2)\hat{\delta}_{ki}^2 ,$$

and $c = n(n-1)(n-2)^{-1}$ with $n = 45$. The coefficient $\hat{\gamma}_k$ is the sample regression coefficient of $\hat{\delta}_{ki}^2$ regressed on δ_{ki}^2 . The

regression estimator v_{k1} and the ratio estimator v_{k2} are constructed using the chosen set of 10 segments. The quantity, $10^{-1} \Sigma_{i=1}^{10} \hat{\delta}_{ki}^2$, is the mean square error of the "Jackknife" estimator, which is constructed based on these 10 segments. The residuals obtained in a regression of ground crop acreage on estimated probability sum using 45 segments serve as auxiliary information in variance estimation.

In Table 4.G.1, we report the variance estimators, v_{k0}, v_{k1}, v_{k2}, for all crops. There is not much difference between the ratio and regression estimator of variance. The usual variance estimator v_{k0} tends to underestimate less than 10 percent for winter wheat, pasture, soybeans, woods, alfalfa, and 'all other'; whereas it tends to underestimate about 20 percent for corn and about 30 percent for sorghum.

Table 4.G.1 - Variance Estimates

Crop	v_{k0}	v_{k1}	v_{k2}	Ratio v_{k2}/v_{k0}
Corn	48.18	60.93	57.08	1.19
Soybeans	37.70	40.92	41.32	1.10
Pasture	76.98	79.12	80.10	1.04
Woods	33.97	37.95	37.80	1.10
Winter Wheat	6.55	6.85	6.85	1.05
Alfalfa	2.84	3.09	3.11	1.10
Sorghum	4.79	6.71	6.23	1.30
All Other	69.78	73.65	73.59	1.05

V. SUMMARY AND CONCLUSIONS

This study investigated alternative methods of using satellite (LANDSAT) data as auxiliary information in the estimation of crop acreages. The LANDSAT data consist of a vector of values for the radiation in four wavelength bands of the electromagnetic spectrum. A LANDSAT scene taken in 1979 in northern Missouri is studied. The eight crops selected for study are corn, winter wheat, pasture, soybeans, woods, alfalfa, sorghum, and all other.

The approach is to use functions of the vector $\underset{\sim}{X}$ as auxiliary variables in regression estimation of the acres in a particular crop. The estimated posterior probability that a point with a satellite value of $\underset{\sim}{X}$ is from crop j was developed as an auxiliary variable. For an individual pixel the true posterior probability is the best possible auxiliary variable because it is the expected value conditional on $\underset{\sim}{X}$ of the crop indicator variable. Based on the estimated posterior probability, a "classification rule" G* was constructed as another auxiliary variable. Let $(p_{1t}, p_{2t}, \ldots, p_{8t})$ be the vector of estimated posterior probabilities for the t-th observation and let $p_{(m)t}$ be the maximum of the eight probabilities. Let $R_{jt} = p_{(m)t}^{-1} p_{jt}$. Let f_j be the fraction of the acres that are in crop j and let $R_{j(f)}$ be the value such that f_j of the R_{jt} values exceed $R_{j(f)}$. Then, the G*-variable for crop j is defined by

$$G^*_{jt} = 1 \text{ , if } R_{jt} = 1$$

$$= 1 \ , \quad \text{if} \quad R_{jt} > R_{j(f)}$$

$$= 0 \ , \quad \text{otherwise} \ .$$

The G*-rule could be optimized by choosing a cutoff point other than $R_{j(f)}$, but no attempt was made to construct such an optimum.

A third auxiliary variable is the USDA classifier. Using the procedure described in Sigman et al. (1978), a crop code is assigned to each pixel. Thus,

$$U_j = 1 \quad \text{if pixel is coded as crop} \quad j$$

$$= 0 \ , \ \text{otherwise} \ .$$

The USDA only assigned codes for the four crops; corn, soybeans, pasture and woods.

The effect of estimating the posterior probability on the variance of the resulting regression estimator for large samples was investigated. An approach we adopted, and that followed by Fuller (1975), is to specify a sequence of finite populations and samples from these populations. The finite population is a random sample from a multivariate infinite population with certain moment conditions. Under certain mild assumptions, the asymptotic normality of the regression estimator of the finite population mean was derived as both the sample

size and population size became large. It was found that the effect of estimating the posterior probability when constructing a regression estimator of the finite population mean is negligible in large samples. However, in estimating the finite population regression coefficient, the error due to estimating auxiliary variates is the same as the error of the least squares estimator for the case with transformed auxiliary variable known.

Two regressions were computed to evaluate the alternative auxiliary variables. The first regression used the individual pixels as observations. In this case, the dependent variable for crop j is one if the ground truth of the pixel is crop j and is zero otherwise. The independent variables are the values of the auxiliary variables for the pixel. The second regression used June Enumerative Survey segment totals as variables. The dependent variable for crop j is the acres of crop j in the segment and the independent variables are the sum of the values of the auxiliary variables for those pixels in the segment. To make comparisons, we assumed that the effect due to the estimation in the discrete auxiliary variates is negligible for large samples. Under this assumption, either the R^2 values or the residual mean square errors can be used for making comparisons among various auxiliary variables.

Two methods of estimating posterior probability were employed in this study. In the first method, the mean vector and covariance matrix were estimated for each crop from data collected in the June Enumerative

survey. The probabilities were then constructed on the assumption that the vector of readings for each crop is distributed as a multivariate normal vector. Because the normal probability model was not supported by the data plot, a second method of estimating the density was used. In this procedure, the probability density was written as

$$p(X_1, X_2, X_3, X_4) = p(X_4 | X_1, X_2, X_3)p(X_3 | X_1, X_2)$$

$$\times p(X_2 | X_1)p(X_1) \ .$$

Each conditional density was assumed to be normal with conditional mean that could be quadratic in the conditioning variables. In this procedure, the conditional variance can also be a function of the conditioning variables, but in our example the conditional variances seemed nearly constant. This method of conditional fitting permits graphical inspection of the fit at each step of the process. Although the multivariate normal model is rejected by the data, the use of the alternative model to construct probabilities made very little improvement in the performance of the regressions. However, the fact that the improvement is uniform indicates that the method of density estimation may sometimes play an important role.

Two methods of modifying density estimates for skewness were considered, transformations and screening, with the screening results presented in this report. Screening deviate observations from the data

set before computing the means and covariances made modest improvement in the performance of the auxiliary variables.

At the pixel level, the posterior probability performs uniformly better than the USDA pixel classifier and uniformly better than the maximum probability classification procedure G*. The R^2 values for the pixel regressions on the posterior probability ranged from 0.01 (for the rare crops alfalfa and sorghum) to 0.43 for soybeans. For corn, soybeans, pasture, and woods the R^2 for the USDA classifier averaged about three fourths of the R^2 for the posterior probability. The G* rule was superior to the USDA classifier for the four crops for which the USDA rule was available, but the differences were not large.

The superiority of the posterior probability as an auxiliary variable largely disappears in the segment level regression. The G*-rule performs as well or better than the probability sum in the segment regressions for all crops except 'all other'. For woods, the mean square error of the G*-rule is about 83 percent of that for the probability sum. For 'all other', the residual mean square error for the probability sum is about 75 percent of that for the G*-rule. The heavy clustering of the segment data seems to result in particularly poor performance of the auxiliary variables for the crops which are not easily discriminated from the others. Corn and woods are two crops for which the $\underset{\sim}{X}$-distributions are similar. For corn, significant improvement over the simple segment regression of acres on the auxiliary variable was achieved in several ways. The most effective way was the

use of a multiple regression with both the G*-rule for corn and the corn probability sum as independent variables. The multiple regression produces a residual mean square error that is about 60 percent of the residual mean square of the G*-rule and about 50 percent of the mean square error of the probability sum.

The G*-rule is superior to the USDA pixel classifier for corn, pasture and woods and was inferior for soybeans in the segment regressions.

A procedure, referred to as the restricted segment multiple regression, applies generalized least squares to the estimation of the system of the segment regressions for all crops. For all crops, except woods, this procedure gave results similar to those obtained in the multiple regression containing G* and the posterior probability as independent variables. For woods, the residual mean square error for the G*-rule is about 80 percent of that for the restricted segment multiple regression procedure.

The use of a multiple regression with both the probability sum and segment acres as independent variables performed nearly as well as the G*-rule. The inclusion of the segment acres made a significant improvement only for corn. The reduction of the residual mean square error associated with the inclusion of the segment acres is about 20 percent for corn.

A jackknife ratio type estimator of the variance for the regression estimator was constructed based on a randomly chosen set of 10 segments

from the sample. One segment from the sample of 45 segments is deleted and the linear regression estimator of crop acreages is constructed using the remaining 44 segments. This process is repeated 10 times. The usual variance estimator is about a 10 percent underestimate for winter wheat, pasture, soybeans, woods, alfalfa, and 'all other'; and about a 20 - 30 percent underestimate for corn and sorghum.

In summary, the use of the posterior probability as an auxiliary variable was generally marginally superior to the USDA classifier as an auxiliary variable. In the case of corn, the use of the posterior probability and a classification variable in a multiple regression produced a residual mean square error about one half of that attainable with either variable alone. This study illuminates the importance of the intrasegment correlation in determining the performance of auxiliary variables. Because indicator variables constructed using classification rules may have smaller intrasegment correlations than that of the posterior probability, the performance of the sum of the indicator variables in the segment regressions may equal or exceed that of the sum of the posterior probabilities.

VI. BIBLIOGRAPHY

Amis, M. L., R. K. Lennington, M. V. Martin, W. G. McGuire, and S. S. Sen. 1981. Evaluation of large area crop estimation techniques using LANDSAT and ground-derived data. LEMSCO-15763, March 1981.

Anderson, J. A. 1972. Separate sample logistic discrimination. Biometrika 59:19-36.

Anderson, T. W. 1951. Classification by multivariate analysis. Psychometrika 16:31-50.

Anderson, T. W. 1958. An introduction to multivariate statistical analysis. John Wiley & Sons, Inc., New York.

Anderson, T. W. 1973a. Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions. Pages 17-35 in T. Cacoullos, ed. Discriminant Analysis and Applications, Academic Press, Inc., New York.

Anderson, T. W. 1973b. An asymptotic expansion of the distribution of the "Studentized" classification statistic W. Ann. Stat. 1:964-972.

Cacoullos, T. 1973. Discriminant analysis and applications. Academic Press, Inc., New York.

Campbell, C. 1977. Properties of ordinary and weighted least squares estimators for two stage samples. Pages 800-805 in Proceedings of the Social Statistics Section, American Statistical Association, Washington, D.C.

Cochran, W. G. 1942. Sampling theory when the sampling units are of unequal sizes. Journal of the American Statistical Association 37:199-212.

Cochran, W. G. 1977. Sampling techniques. John Wiley & Sons, Inc., New York.

Cox, D. R. 1966. Some procedures associated with the logistic response curve. Research Papers in Statistics: Festschrift for J. Neyman. John Wiley & Sons, Inc., New York.

Craig, M. E., R. S. Sigman, and M. Cardenas. 1978. Area estimation by LANDSAT: Kansas 1976 winter wheat. Economics, Statistics, and Cooperative Service, U.S. Department of Agriculture, Washington, D.C., August 1978.

Day, N. E. and D. F. Kerridge. 1967. A general maximum likelihood discriminant. Biometrics 23:313-323.

DeMets, D. and M. Halperin. 1977. Estimation of a simple regression coefficient in samples arising from a subsampling procedure. Biometrics 33:47-56.

Duda, R. and P. Hart. 1973. Pattern classification and scene analysis. John Wiley & Sons, Inc., New York.

Fisher, R. A. 1936. The use of multiple measurement in taxonomic problems. Annals of Eugenics 7:179-188.

Frankel, M. R. 1971. Inference from survey samples. Ann Arbor: Institute for Social Research, The University of Michigan, Ann Arbor.

Fuller, W. A. 1975. Regression analysis for sample survey. Sankhyā, Series C 37:117-132.

Fuller, W. A. 1976. Introduction to statistical time series. John Wiley & Sons, Inc., New York.

Fuller, W. A. 1977. Personal communication to William Wigton, December 1977. Department of Statistics, Iowa State University, Ames, Iowa.

Fuller, W. A. 1981. Notes on the theory of econometrics for Statistics 538, Department of Statistics, Iowa State University, Ames, Iowa.

Fuller, W. A. and G. E. Battese. 1973. Transformations for estimation of linear models with nested-error structure. Journal of the American Statistical Association 68:626-632.

Gilbert, E. S. 1969. The effect of unequal variance-covariance matrices on Fisher's linear discriminant functions. Biometrics 25:505-515.

Gleason, C., R. R. Starbuck, R. S. Sigman, G. A. Hanuschak, M. E. Craig, P. W. Cook, and R. D. Allen. 1977. The auxiliary use of LANDSAT data in estimating crop acreages: Results of the 1975 Illinois crop-acreage experiment. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., October 1977.

Hand, D. J. 1981. Discrimination and classification. John Wiley & Sons, Inc., New York.

Hanuschak, G. A. and M. Cardenas. 1978. Multiple regression estimation using classified LANDSAT data. Economics, Statistics, and Cooperative Service, U.S. Department of Agriculture, Washington, D.C., April 1978.

Henderson, H. V. and S. R. Searle. 1979. Vec and vech operators for matrices with some uses in Jacobians and multivariate statistics. The Canadian Journal of Statistics 7(1):65-81.

Hidiroglou, M., W. A. Fuller, and R. D. Hickman. 1980. SUPER CARP. Sixth Edition. Survey Section, Statistical Laboratory, Iowa State University, Ames, Iowa.

Hills, M. 1966. Allocation rules and their error rates. Journal of the Royal Statistical Society B28:1-20.

Holt, D. and A. J. Scott. 1981. Regression analysis using survey data. The Statistician 30:169-178.

Holt, D., T. M. F. Smith, and P. D. Winter. 1980. Regression analysis of data from complex surveys. Journal of the Royal Statistical Society A143:474-487.

Kish, L. 1965. Survey sampling. John Wiley & Sons, Inc., New York.

Kish, L. and M. R. Frankel. 1974. Inference from complex samples. Journal of the Royal Statistical Society B36:1-22.

Konijn, H. S. 1962. Regression analysis for sample surveys. Journal of the American Statistical Association 57:590-606.

Lachenbruch, P. A. 1968. On expected values of probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. Biometrics 24:823-834.

Lachenbruch, P. A. and M. R. Mickey. 1968. Estimation of error rates in discriminant analysis. Technometrics 10:1-11.

Lennington, R. K. and H. Malek. 1978. The CLASSY clustering algorithm-description, evaluation, and comparison with the iterative self-organizing clustering system (ISOCLS). LEC-11289, March 1978.

Lennington, R. K. and M. E. Rassbach. 1978. An adaptive maximum likelihood clustering algorithm. LEC-12145, May 1978. Presented at the Ninth Annual Meeting of the Classification Society (North American Branch), Clemson University (Clemson, South Carolina), May 21-23, 1978.

Lennington, R. K. and M. E. Rassbach. 1979a. Mathematical description and program documentation for CLASSY, an adaptive maximum likelihood clustering method. LEC-12177 (JSC-14621), April 1979.

Lennington, R. K. and M. E. Rassbach. 1979b. ClASSY - An adaptive maximum likelihood clustering algorithm. Proceedings of Technical Sessions, Volume II, LACIE Symposium, October 1978, JSC-16015 (Houston, Texas), July 1979.

Marks, S. and O. J. Dunn. 1974. Discriminant functions when covariance matrices are unequal. Journal of the American Statistical Association 69:555-559.

McLachlan, G. J. 1974a. An asymptotic unbiased technique for estimating the error rates in discriminant analysis. Biometrics 30:239-249.

McLachlan, G. J. 1974b. Estimation of the errors of misclassification on the criterion of asymptotic mean square error. Technometrics 16:255-260.

McLachlan, G. J. 1974c. The relationship in terms of asymptotic mean square error between the separate problems of estimating each of the three types of error rates of the linear discriminant function. Technometrics 16:569-575.

McLachlan, G. J. 1975. Confidence intervals for the conditional probability of misallocation in discriminant analysis. Biometrics 31:161-167.

McLachlan, G. J. 1976. The bias of the apparent error rate in discriminant analysis. Biometrika 63:239-244.

Mergerson, J. W. 1981. Crop area estimates using ground-gathered and LANDSAT data: A multiemporal approach, Missouri 1979. Statistical Research Division, Economics and Statistics Service, U.S. Department of Agriculture, Washington, D.C., February 1981.

Nathan, G. and D. Holt. 1980. The effect of survey design on regression analysis. Journal of the Royal Statistical Society B42:377-386.

Neudecker, H. 1969. Some theorems on matrix differentiations with special reference to Kronecker matrix products. Journal of the American Statistical Association 64:953-963.

Okamoto, M. 1963. An asymptotic expansion for the distribution of the linear discriminant function. Annals of Mathematical Statistics 34:1286-1301.

Rao, C. R. 1952. Advanced statistical methods in biometric research. John Wiley, Inc., New York.

Rao, C. R. 1965. Linear statistical inference and its applications. John Wiley & Sons, Inc., New York.

Scott, A. J. and D. Holt. 1982. The effect of two-stage sampling on ordinary least squares methods. Journal of the American Statistical Association 77:848-854.

Sigman, R. S., C. P. Gleason, G. A. Hanuschak, and R. A. Starbuck. 1977. Stratified acreage estimates in the Illinois crop-acreage experiment. Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

Sigman, R. S., G. A. Hanuschak, M. E. Craig, P. W. Cook, and M. Cardenas. 1978. The use of regression estimation with LANDSAT and probability ground sample data. Presented at 1978 Annual Meeting of the American Statistical Association, San Diego, California, August 1978.

Statistical Analysis System. 1982. SAS user's guide: Statistics. SAS Institute Inc., Raleigh, North Carolina.

Swain, P. H. and S. M. Davis. 1978. Remote sensing: The quantitative approach. McGraw-Hill Book Company, New York.

Toussaint, G. T. 1974. Bibliography on estimation of misclassification. IEEE Transactions on Information Theory IT-20:472-479.

Wahl, P. W. and R. A. Kronmal. 1977. Discriminant functions when covariances are unequal and sample sizes are moderate. Biometrics 33:479-484.

Wald, A. 1944. On a statistical problem arising in the classification of an individual into one of two groups. Annals of Mathematical Statistics 15:145-162.

Zeller, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association 57:348-368.

VII. ACKNOWLEDGEMENTS

I acknowledge with thanksgiving that God grants success. I thank Him for the strength and stability of his presence in me.

I wish to express my sincere appreciation and gratitude to Distinguished Professor Wayne A. Fuller, who was a constant source of expert guidance during the course of this study. Without his assistance and patience, I certainly would not have been able to complete this work.

I thank Jane Stowe for her efficient and excellent typing of this manuscript. Thanks also go to Carol Francisco, who has provided excellent comments. And also, I am grateful to Helen Nelson for her assistance and friendship.

This work was supported by the U.S. Department of Agriculture under the research agreement 58-319T-1-0054X.